A STEPPING PROCEDURE BASED ON A LOCAL INFLUENCE MEASURE TO IDENTIFY MULTIPLE MULTIVARIATE OUTLIERS

TSE SUK YAN

A Thesis Submitted in Partial Fulfillment

of the Requirements for the Degree of

Master of Philosophy

in

Statistics

June 2000


香港中文大學
圖書館系統
22 JUN 2001
UNIVERSITY
LIBRARY SYSTEM
THE CHINESE UNIVERSITY OF HONG KONG

THE CHINESE UNIVERSITY OF HONG KONG

GRADUATE SCHOOL

The undersigned certify that we have read a thesis, entitled "A Stepping Procedure Based on a Local Influence Measure to Identify Multiple Multivariate Outliers" submitted to the Graduate School by Suk-Yan Tse (謝淑茵) in partial fulfillment of the requirements for the degree of Master of Philosophy in Statistics. We recommend that it be accepted.

Dr. W.Y. Poon,

Supervisor

Prof. S.Y. Lee

Prof. T.S. Lau

Prof. C.P. Chou,

External Examiner

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institution of learning.

Acknowledgment

I would like to express my deepest thanks to my supervisor, Dr. Wai-Yin Poon, for her encouragement and guidance during the period of this research program. I would also like to thank Mr. K. H. Leung for his technical consultation. It is also a pleasure to express my gratitude to all the staff of the Department of Statistics for their kind assistance.

Abstract

In this thesis, a new stepping procedure based on a local influence measure to identify multiple multivariate outliers is proposed. The new procedure is based on the stepping procedure given in Hadi (1992) and the outlier measure and the benchmark given in Poon, Lew & Poon (2000). Two kinds of metrics, the sample covariance matrix and the sample robust covariance matrix, are considered. Moreover, two kinds of methods for developing the stopping criterion are examined. The first method uses only observations in the basic subset and the second method uses all the observations in the data set. Reported data sets from the literature and artificial data sets are used for illustration and simulation studies are performed to investigate the performance of the new procedure. It can be shown that the second method accompanied with either one of the two metrics gives better results while the first method gives unsatisfactory results. The stepping procedure involves a selection of so-called initial basic subset by certain criteria and a modification of the procedure motivated by Atkinson (1994) that selects initial basic subset randomly is considered. Initial basic subset selected randomly is likely to contain outliers but results on analyzing several data sets by the modified procedure indicate that whether the initial subsets contain outliers or not does not affect the performance of the proposed procedure. Finally, some aspects of the mentioned procedures are discussed, including a modification which improves the first method.

摘要

在這篇論文中，我們提出一個以局部影響度量爲基礎的新步驟程序來鑑定複合多元離群值。這個新程序是建於 Hadi (1992) 所提出的步驟法以及 Poon, Lew & Poon (2000) 所提出的離群測量值和基準。我們會考慮兩種度量和兩種用作發展中止準則的方法，這兩種度量包括樣本協方差矩陣和樣本穩健協方差矩陣，而發展中止準則的第一種方法是使用基本子集中的觀測值，第二種方法是採用整個數據集的觀測值。其他論文中的報告數據集和人造數據集會用作舉例說明這個新程序，而模擬研究則用作審查新程序的表現，結果指出第二種方法配合以上任意一種度量是比較優勝，而第一種方法所得出的結果就不令人滿意。這個新步驟程序包括根據某些基準來選取原始基本子集，而我們會根據 Atkinson (1994) 這篇論文來修改新步驟程序，使原始基本子集可以隨機地選取。隨機地選取的原始基本子集可能包含離群值，不過，採用修改後的程序以及分析了多個數據集所得出的結果顯示原始子集包含離群值與否是不會影響新步驟程序的表現。最後，我們會討論以上提及過的程序的幾方面，包括提出一個修改來改良第一種方法。

Contents

		Page
Chapter 1	**Introduction**	1
Chapter 2	**The Elements of the New Procedure**	6
2.1	The Stepping Algorithm	6
2.2	Outlier Measure and Benchmark	17
Chapter 3	**The New Procedure**	21
3.1	Procedure	22
3.2	Examples	31
3.3	Simulation Study	41
	3.3.1 Terms and Factors	41
	3.3.2 Procedure	45
	3.3.3 Results	47
Chapter 4	**Robust Version of the New Procedure**	51
4.1	Procedure	52
4.2	Examples	58
4.3	Simulation Study	62
	4.3.1 Procedure	63
	4.3.2 Results	65
Chapter 5	**The New Procedure with Random Initial Subset**	68
5.1	The Elements	69
5.2	Procedure	71
5.3	Examples	77
Chapter 6	**Discussion**	91
Chapter 7	**Conclusion**	102
Appendix		105
References		114

Chapter 1

Introduction

In the literature, Mahalanobis distance can be used to distinguish between good observations and outliers in multivariate data. It measures the distance from each observation to the center of the cloud of observations or the sample mean, relative to the sample covariance matrix. Good observations are the majority of data coming from a well-behaved population. Outliers are observations that are far away from the majority of data. Outliers may be shown by the large values of Mahalanobis distances. The Mahalanobis distance is a powerful tool of detecting outliers when there is only one outlier. However, Mahalanobis distance is not effective in identifying outliers when the number of outliers increases. Masking may occur when the number of outliers is greater than one. Several outliers that form a point cloud may attract the sample mean towards the point cloud and distort the sample covariance matrix, causing the Mahalanobis distance to be small. Multivariate outliers may be masked and may not be detected as outliers while other good observations may be treated as outliers instead. Therefore, robust

estimates of location and covariance matrix are used to replace the sample mean and the sample covariance matrix in the Mahalanobis distance in the literature.

The methods of detecting multivariate outliers and regression outliers are similar. The Mahalanobis distance or the robust distance is used to identify multivariate outliers and the least median of squares (LMS) is used to detect regression outliers. Cutoff values are used in both methods to determine which observations are the outliers. Here are some methods for detecting multivariate outliers by using the Mahalanobis distance or the robust distance and for identifying regression outliers by using the least median of squares.

Atkinson (1986) uses a two-stage method, including the exploratory stage and the confirmatory stage, to detect outliers. In the exploratory stage, elemental sets of p observations are sampled repeatedly from the n observations to perform least median of squares regression until a clear pattern of outliers is seen, where p is the rank of the regression model. In the confirmatory stage, some efficient diagnostic methods of least squares regression based on the deletion of the potential outliers identified in the exploratory stage are used to confirm the presence of outliers.

Rousseeuw & van Zomeren (1990) proposed a highly robust regression method by least median of squares to detect outliers. A plot of the robust residuals against the robust distances is used to find out the leverage points and the regression outliers.

Hadi (1992) proposed a procedure for identifying multiple outliers in multivariate data. In the first step, i.e. Step 0 in his procedure, the data are ordered according to a robust measure. The ordered data are then divided into two initial subsets called a basic subset and a non-basic subset. In the second and third steps, i.e. Steps 1 and 2 in his procedure, the observations are rearranged in ascending order according to a distance measure which measures the distance from each data point to the mean of the basic subset, relative to the covariance matrix of the basic subset. The basic subset and the non-basic subset are updated by increasing the number of observations in the basic subset by one and decreasing the number of observations in the non-basic subset by one according to the ordered observations obtained above. These two steps are repeated until a stopping criterion is met, for example, when the number of observations in the basic subset m equals the integer part of $(n + p + 1)/2$ where n is the sample size and p is the dimension of the data. The observations in the final non-basic subset are declared as the outliers.

Atkinson & Mulira (1993) considered a forward procedure using Mahalanobis distance with a random initial subset. The procedure is similar to that proposed by Hadi (1992). However, Atkinson & Mulira's (1993) procedure stops when the number of observations m used to calculate the mean and the covariance matrix equals the total number of observations n while Hadi's (1992) procedure stops when m equals the integer part of $(n + p + 1)/2$ which is less than n. Atkinson

& Mulira (1993) also proposed to use stalactite plot and index plot to provide a clear view of the outliers.

Atkinson (1994) proposed two algorithms for detecting outliers. One uses the least median of squares for regression model and the other uses the Mahalanobis distance and minimum volume ellipsoid (MVE) for multivariate data. The algorithm for identifying multivariate outliers is similar to Hadi's (1992). However, the initial subset in the algorithm of Atkinson (1994) is selected randomly. That is, the calculations of robust estimates of location and dispersion at the beginning of the algorithm are not needed. The values of minimum volume ellipsoid are recorded and the patterns of outliers are shown in stalactite plots. In order to find the most suitable basic subset for identifying outliers precisely, the most appropriate basic subset is chosen such that it gives the smallest value of minimum volume ellipsoid. The algorithm for detecting regression outliers resembles the one for detecting multivariate outliers but the least median of squares is used.

For all the procedures as mentioned above, distributional assumption is needed. The squared Mahalanobis distance and the squared robust distance follow a chi-squared distribution with p degrees of freedom when the sample size is large, where p is the dimension of the data set. The cutoff point of determining whether an observation is an outlier or not is also based on a chi-squared distribution approximately. However, the distribution of the distances is difficult to obtain when the sample size is not large enough.

In this thesis, a new stepping procedure of identifying outliers in multivariate case based on a measure developed from the local influence approach, which is distributional assumption free, is proposed. The new procedure is based on the stepping algorithm of Hadi (1992) and the outlier measure together with the benchmark proposed by Poon, Lew & Poon (2000). The outlier measure and the benchmark are based on the local influence approach of Cook (1986). The particulars of using the stepping algorithm of Hadi (1992) as well as the outlier measure and the benchmark of Poon, Lew & Poon (2000) are given in the next chapter.

Here is the framework of this thesis. The details of Hadi's (1992) stepping algorithm and Poon, Lew & Poon's (2000) outlier measure together with its benchmark are described in Chapter 2. The new stepping procedure of detecting multivariate outliers is proposed in Chapter 3 and some reported data sets and simulation studies are used to illustrate the procedure. A refinement of the new procedure is given in Chapter 4. The refined procedure is based on the robust version of the outlier measure and the benchmark. Reported data sets and simulation are also applied to the revised version of the procedure. In Chapter 5, a procedure based on randomly chosen initial subsets which may contain outliers is considered. A discussion and a conclusion are given in Chapter 6 and Chapter 7 respectively.

Chapter 2

The Elements of the New Procedure

Hadi's (1992) stepping algorithm and Poon, Lew & Poon's (2000) outlier measure are the basic elements of the new procedure proposed in this thesis. These elements are described in the following and the particulars of using them are also given. The new procedure based on these elements is given in the next chapter.

2.1 The Stepping Algorithm

Hadi's (1992) stepping algorithm for identifying multiple outliers in multivariate data is as follows:

Step 0: Initial Ordering

1. Define $\mathbf{X_{n\times p}}$ be the data set with n observations and dimension p, that is

$$\mathbf{X_{n\times p}} = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix},$$

where x_{ij} is the ith observation of the jth dimension, $i = 1, \dots, n$, $j = 1, \dots, p$ and

$$x_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix} \tag{2.1}$$

is the vector containing the ith elements.

2. Initially rearrange the n observations in ascending order according to a robust distance,

$$D_i(C_R, S_R) = \sqrt{(x_i - C_R)^T S_R^{-1} (x_i - C_R)}\,, \qquad i = 1, \dots, n, \tag{2.2}$$

where C_R and S_R are robust location and covariance matrix estimators obtaining in the following.

   (a) Compute the co-ordinatewise median vector

$$C_M = \begin{pmatrix} C_{M1} \\ \vdots \\ C_{Mp} \end{pmatrix},$$

where C_{Mj}, $j = 1, \ldots, p$, is the median of the elements in

$$\begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix}.$$

(b) Compute

$$S_M = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - C_M)(x_i - C_M)^T$$

which is a robust version of the sample covariance matrix.

(c) Rearrange the observations in ascending order according to the robust distance $D_i(C_M, S_M)$ which measures the distance of the ith observation from the co-ordinatewise median vector, relative to the robust covariance matrix, where

$$D_i(C_M, S_M) = \sqrt{(x_i - C_M)^T S_M^{-1} (x_i - C_M)}\,, \qquad i = 1, \ldots, n.$$

(d) Define the weight function

$$v_i = \begin{cases} 1, & \text{if } i \leq \text{ integer part of } (n+p+1)/2, \\ 0, & \text{otherwise,} \end{cases} \qquad i = 1, \ldots, n.$$

Observations that are close to the co-ordinatewise median vector of the data relative to the robust covariance matrix are given larger weights. The weights for the first integer part of $(n+p+1)/2$ ordered observations with smaller robust distances $D_i(C_M, S_M)$ are one. The weights for the observations with larger robust distances are zero as the observations are far away from the co-ordinatewise median vector of the data.

(e) Compute C_R the robust location and S_R the covariance matrix estimators by using the weight function and the ordered observations, where

$$C_R = \frac{\sum_{i=1}^{n} v_i x_i}{\sum_{i=1}^{n} v_i} \quad \text{and} \quad S_R = \frac{\sum_{i=1}^{n} v_i (x_i - C_R)(x_i - C_R)^T}{\sum_{i=1}^{n} v_i - 1}.$$

The observations that are close to the co-ordinatewise median vector are used to compute C_R and S_R. This set of observations is unlikely to contain outliers.

3. Divide the observations into two initial subsets:

   - a basic subset containing the first $p+1$ observations, and
   - a non-basic subset containing the last $n-p-1$ observations.

The observations are rearranged in ascending order using the robust distance $D_i(C_R, S_R)$ (equation (2.2)). Outliers are observations with large robust distances. Large robust distances indicate that the observations are

far away from the subset that contains no outliers. The basic subset is intended to be outlier free. So, observations with small robust distances form the basic subset while those with large robust distances form the non-basic subset.

4. Go to Step 3 and check whether the stopping criterion is met before go to Steps 1 and 2.

Step 1 a): Basic Subset of Full Rank

If the basic subset is of full rank, the corresponding covariance matrix S_b is non-singular.

Compute

$$\sqrt{(x_i - C_b)^T S_b^{-1} (x_i - C_b)}\ , \qquad i = 1, \ldots, n, \tag{2.3}$$

where C_b and S_b are the mean and covariance matrix of the basic subset.

Step 1 b): Basic Subset Not of Full Rank

1. If the basic subset is not of full rank, the eigenvalues of S_b, $\lambda_1 \geq \ldots \geq \lambda_p = 0$, and the matrix V_b containing the corresponding set of normalized eigenvectors are computed.

2. Calculate

$$\sqrt{(x_i - C_b)^T V_b W_b V_b^T (x_i - C_b)}\ , \qquad i = 1, \ldots, n, \tag{2.4}$$

where W_b is a diagonal matrix whose jth diagonal element is

$$w_j = \frac{1}{\max\{\lambda_j, \lambda_s\}}, \qquad j = 1, \ldots, p \tag{2.5}$$

and λ_s is the smallest non-zero eigenvalue of S_b.

3. The reason why the eigenvalues and eigenvectors are used is given by Hadi (1992) as below:

If the basic subset is not of full rank, S_b is singular and the inverse of S_b does not exist. Therefore, equation (2.3) cannot be used. However, eigenvalues and eigenvectors of S_b can be used to compute the inverse of a singular covariance matrix as in the following.

(a) Express S_b as

$$S_b = V_b \Lambda_b V_b^T,$$

where

$$\Lambda_b = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix}$$

is a diagonal matrix containing the eigenvalues of S_b, $\lambda_1 \geq \ldots \geq \lambda_p = 0$ and V_b is the matrix containing the corresponding set of normalized eigenvectors.

(b) Note that when S_b is non-singular, equation (2.3) can be expressed as

$$\sqrt{(x_i - C_b)^T V_b \Lambda_b^{-1} V_b^T (x_i - C_b)}\,, \qquad i = 1, \ldots, n. \tag{2.6}$$

(c) When S_b is singular and $\lambda_p = 0$, $1/\lambda_p$ does not exist, and thus

$$\Lambda_b^{-1} = \begin{pmatrix} \frac{1}{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\lambda_p} \end{pmatrix}$$

does not exist.

Therefore, the diagonal elements $1/\lambda_j$ of Λ_b^{-1}, $j = 1, \ldots, p$, are modified as in equation (2.5) and Λ_b^{-1} is replaced by W_b as in equation (2.4). From equation (2.5), the smallest non-zero eigenvalue of S_b is chosen to calculate the jth diagonal element if the jth eigenvalue is not positive.

When S_b is non-singular, equations (2.3), (2.4), (2.6) are equivalent; however, equation (2.3) is preferred as the computations of the eigenvalues and eigenvectors of S_b are not needed.

Step 2: Increase Size of Basic Subset

1. Rearrange the observations in ascending order according to either expression (2.3) or (2.4) depending on whether S_b is of full rank or not.

2. Let r be the number of observations in the current basic subset.

   Divide the observations into two subsets:

   - a basic subset containing the first $r+1$ observations, and
   - another subset containing the remaining $n-r-1$ observations.

   The number of observations in the basic subset is increased by one while that in the non-basic subset is decreased by one according to the ordered observations obtained by the distance measure in equation (2.3) or (2.4). Steps 1 and 2 are repeated until a stopping criterion in the next step is met.

Step 3: Stopping Criterion

1. Compute the robust distance

$$D_i(C_b, S_b) = \sqrt{(x_i - C_b)^T (c_b S_b)^{-1} (x_i - C_b)}\ , \qquad i = 1, \ldots, n,$$

   where

$$c_b = c_{npr} m_j / \chi^2_{p,0.50},$$

   according to Hadi (1992), is a correction factor when the data come from a multivariate normal distribution. m_j is the $100(h/n)$th percentile of the n values in equation (2.3) or (2.4) depending on whether the basic subset is

of full rank or not where h is the integer part of $(n+p+1)/2$ and number of good points should be at least h from Lopuhaä & Rousseeuw (1991). $\chi^2_{p,0.5}$ is the 0.5 probability point of the chi-squared distribution with p degrees of freedom. An appropriate small sample correction factor is

$$c_{npr} = \left\{1 + \frac{r}{n-p}\right\}^2,$$

where r is the number of observations in the final basic subset.

Hadi (1994) suggests another correction factor c_{np} with a modification of Hadi's (1992) stepping algorithm where

$$c_{np} = \left(1 + \frac{p+1}{n-p} + \frac{1}{n-h-p}\right)^2 = \left(1 + \frac{2}{n-1-3p} + \frac{p+1}{n-p}\right)^2$$

and h is defined as before.

The correction factor depends only on the sample size n and dimension p but not the number of observations r in the final basic subset. Hadi (1994) compares the performance of the two algorithms by a simulation study when dimension p equals 2 or 5. When $p = 2$, the algorithm of Hadi (1994) is superior to that of Hadi (1992) and is less influenced by masking effect. However, the algorithm of Hadi (1992) performs better when $p = 5$. As the two algorithms are not compared at other values of p, it cannot be

confirmed that the modified algorithm also works well at other values of p. The original algorithm is adopted instead of the modified one to form the new procedure as $p = 5$ or larger values of p will be used in the simulation studies, which will give a clear view of the new procedure, in later chapters.

2. Repeat Steps 1 and 2 until a certain stopping criterion is met. Two stopping criteria are

   (a) Stop when

   $$\min\{D_i(C_b, S_b); i \in \text{non-basic subset}\} \geq C_\alpha$$

   where the critical value C_α is chosen such that

   $$\Pr[\min\{D_i(C_b, S_b); i \in \text{non-basic subset}\} \geq C_\alpha | \mathbf{X} \text{ contains no outliers}] = 1 - \alpha$$

   where $\mathbf{X}$ is a $n \times p$ matrix representing a random sample of size n from a p-dimensional population.

   However, the distribution of $D_i(C_b, S_b)$ is difficult to derive and so is C_α. This stopping criterion can only be used when C_α is known; otherwise, the next stopping criterion is adopted instead.

   (b) Stop when the basic subset contains h observations where h is the integer part of $(n + p + 1)/2$.

Particulars

Compare to other procedures avaliable in the literature, Hadi's (1992) procedure is computationally inexpensive. Most of the procedures require resampling to find the most suitable initial subset; however, only one initial subset of observations that contains no outliers is needed to calculate the robust location and covariance matrix estimates in the robust distance in the initial step of Hadi's (1992) procedure. For other procedures such as Atkinson (1986) and Atkinson (1994), the initial subsets are found by resampling for many times. And a certain criterion is used to determine the most appropriate basic subset as described in the procedure of Atkinson (1994). Hadi's (1992) procedure saves the time of finding the initial subset.

Moreover, Hadi's (1992) procedure takes into account the fact that observations in the basic subset may be dependent and hence produce estimate of covariance matrix that is not of full rank. If the covariance matrix is not of full rank, i.e. singular, the inverse of the covariance matrix does not exist and the Mahalanobis distance or other robust distances that involve the inverse of the covariance matrix cannot be computed. Some authors, for example Rousseeuw & van Zomeren (1990), choose to ignore those initial subsets with singular covariance matrix. Hadi (1992) addressed the problem by replacing the inverse of the covariance matrix by using the eigenvalues and eigenvectors of the covariance matrix as mentioned in equations (2.4) and (2.5).

2.2 Outlier Measure and Benchmark

The outlier measure and the benchmark from Poon, Lew & Poon (2000) are adopted in the new procedure. The outlier measure is developed by using Cook's (1986) local influence approach and its recent modification by Poon & Poon (1999).

Local Influence Approach

The local influence approach of Cook (1986) is given as below:

Let $L(\theta)$ and $L(\theta|w)$ be the log-likelihoods for a postulated model and a perturbed model, where θ is a $p \times 1$ vector of unknown parameters, $w = (w_1, \ldots, w_n)^T$ is a $n \times 1$ vector in Ω of $\mathbf{R}^n$ and Ω represents the set of relevant perturbations.

Assume that there is an w_0 such that for all θ,

$$L(\theta) = L(\theta|w_0).$$

Let $\hat{\theta}$ and $\hat{\theta}_w$ be the maximum likelihood estimator under $L(\theta|w_0)$ and $L(\theta|w)$ respectively. The likelihood displacement defined by Cook (1986) is

$$f(w) = 2(L(\hat{\theta}|w_0) - L(\hat{\theta}_w|w_0)). \tag{2.7}$$

Cook (1986) uses the normal curvature C_l of the graph of the likelihood displacement function (2.7) along a direction l at the optimal point w_0 to study

characteristics of the influence graph. The local influence is strong if the value of C_l is large.

Outlier Measure and its Benchmark

Poon, Lew & Poon (2000) use the conformal normal curvature developed from the normal curvature as the outlier measure. The conformal normal curvature B_j of Poon & Poon (1999) is a one–one function of the normal curvature C_l where B_j is defined in the following:

Let V be a $p \times p$ positive definite matrix representing a chosen known metric, the location estimate $\hat{\mu}$ of μ relative to the chosen metric V is obtained by maximizing the function

$$L(\mu) = -\sum_{i=1}^{n}(x_i - \mu)^T V(x_i - \mu). \tag{2.8}$$

Consider the case-weights perturbation given by

$$L(\mu|w) = -\sum_{i=1}^{n} w_i(x_i - \mu)^T V(x_i - \mu), \tag{2.9}$$

where $w = (w_1, \ldots, w_n)^T$.

Let $\hat{\mu}$ and $\hat{\mu}_w$ be the maximum likelihood estimators under (2.8) and (2.9) respectively. If an observation is outlying in location, its influence on $\hat{\mu}$ is large. A

displacement $f(w)$ is used to assess the influence of an observation to $\hat{\mu}$, where

$$f(w) = 2(L(\hat{\mu}|w_0) - L(\hat{\mu}_w|w_0)). \tag{2.10}$$

Similar to the normal curvature of Cook (1986), the conformal normal curvature B_j for the displacement function (2.10) is used to assess the influence of an observation j. Strong local influence is indicated by the large value of B_j, where

$$B_j = \frac{(x_j - \hat{\mu})^T V (x_j - \hat{\mu})}{\sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n}\left[(x_k - \hat{\mu})^T V (x_l - \hat{\mu})\right]^2}}, \qquad j = 1, \ldots, n, \tag{2.11}$$

where $\hat{\mu}$ is the estimate of the location parameter, V is a chosen known metric, and x_i, x_j, x_k and x_l are vectors of the form (2.1).

Poon, Lew & Poon (2000) proposed that the largeness of B_j can be assessed by a benchmark $2b$, where

$$b = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu})^T V (x_i - \hat{\mu})}{n\sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n}\left[(x_k - \hat{\mu})^T V (x_l - \hat{\mu})\right]^2}}, \qquad i = 1, \ldots, n.$$

Since b is the value of B_j when the contribution of all B_j's is uniform, Poon, Lew & Poon (2000) consider those observations with B_j values greater than the benchmark $2b$ as outliers.

Particulars

The outlier measure and the benchmark proposed by Poon, Lew & Poon (2000) which will be used to develop a new stepping procedure have several nice properties. As the outlier measure and the benchmark are formed by using geometrical concept, no distributional assumption on $\mathbf{X}$ is needed. The measure assumes value between zero and one and it is easier to interpret its magnitude. Different metrics V, including robust estimate, can be used to compute B_j as shown in Poon, Lew & Poon (2000). When the metric V is chosen appropriately, the outlier measure is closely related to the Mahalanobis distance. Specifically, the outlier measure is the Mahalanobis distance divided by a constant (see equation (2.11)).

Chapter 3

The New Procedure

In this chapter, a new procedure of identifying multiple outliers in multivariate data based on the stepping algorithm of Hadi (1992) and the outlier measure B_j and the benchmark $2b$ from Poon, Lew & Poon (2000) is proposed. The new procedure inherits the nice properties of the procedures in Hadi (1992) and in Poon, Lew & Poon (2000) so that it is computationally inexpensive and geometrically orientated. It also has the flexibility of using different metrics and the practicability of handling singular covariance matrix. It is computationally inexpensive as Hadi's (1992) stepping algorithm, which is computationally inexpensive as explained in Chapter 2, is adopted. Moreover, singular covariance matrix can be handled by using Hadi's (1992) stepping algorithm as stated in Chapter 2. The outlier measure and the benchmark of Poon, Lew & Poon (2000) are used to make the new procedure geometrically orientated and has the flexibility of using different metrics. Both the outlier measure and the benchmark are geometrically orientated as described in Chapter 2. That is, no distributional assumption is

needed and whether the sample size is large or not will not affect the validity of the new procedure. Different metrics can be applied in the outlier measure, including the sample covariance matrix and the robust estimate of the covariance matrix, so that the new procedure unifies many other procedures proposed in the literature.

The new procedure is given in the first section of this chapter. The notation B_i is used for the proposed procedure instead of B_j from this chapter onwards. Some reported data sets are applied to the proposed procedure to see the effectiveness of the procedure in the second section. Results of simulation study are presented in the third section to give a throughout picture of the performance of the procedure.

3.1 Procedure

The new procedure is similar to that of Hadi's (1992). One of the differences is that the robust distance and the distance measure are replaced by the outlier measure proposed by Poon, Lew & Poon (2000). The other difference is that the stopping criterion of the new procedure is not distributionally orientated but geometrically orientated. The benchmark $2b$ proposed by Poon, Lew & Poon (2000) is used in the stopping criterion. The new procedure is proposed as follows:

Step 0: Initial Ordering

1. Rearrange the n observations in ascending order according to an outlier measure B_i, where

$$B_i = \frac{(x_i - \hat{\mu})^T S^{-1}(x_i - \hat{\mu})}{\sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n}\left[(x_k - \hat{\mu})^T S^{-1}(x_l - \hat{\mu})\right]^2}}, \qquad i = 1, \ldots, n$$

where S is the sample covariance matrix which is symmetric and positive definite.

$$\hat{\mu} = \begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_p \end{pmatrix}, \qquad j = 1, \ldots, p$$

is the estimate of the location parameter, where $\hat{\mu}_j$ is the mean of the elements in the jth column vector

$$\begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix}, \qquad j = 1, \ldots, p$$

of $\mathbf{X_{n\times p}}$, and

$$\mathbf{X_{n\times p}} = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} = \begin{pmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \ldots & x_{np} \end{pmatrix}$$

is a $n \times p$ matrix representing a random sample of size n from a p-dimensional population and x_i is the column vector

$$\begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix},$$

where x_{ij} is the ith observation of the jth dimension where $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

2. After computing B_i, divide the observations into two initial subsets according to the ascending order of the n observations:

   - a basic subset containing the first $p + 1$ observations, and
   - a non-basic subset containing the last $n - p - 1$ observations.

3. Go to Step 3 and check the stopping criterion before go to Steps 1 and 2.

Step 1: Use Basic Subset

In Step 1, the mean C_r and the sample covariance matrix S_r of the basic subset replace the mean $\hat{\mu}$ and the sample covariance matrix S of the whole data set respectively in calculating B_i in Step 0. Two cases, depending on whether the sample covariance matrix of the basic subset is of full rank or not, are considered in calculating the outlier measure B_i (see equations (3.1) and (3.2) respectively).

Case 1: Basic Subset of Full Rank

If the sample covariance matrix S_r of the basic subset is of full rank, compute

$$B_i(full) = \frac{(x_i - C_r)^T S_r^{-1}(x_i - C_r)}{\sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n}\left[(x_k - C_r)^T S_r^{-1}(x_l - C_r)\right]^2}}, \qquad i = 1, \ldots, n, \tag{3.1}$$

where C_r and S_r are the mean and covariance matrix of the basic subset.

Case 2: Basic Subset Not of Full Rank

1. If S_r is not of full rank, compute the eigenvalues of S_r, $\lambda_{r1} \geq \ldots \geq \lambda_{rp} = 0$, and the corresponding normalized eigenvectors matrix V_r.

2. Compute

$$B_i(not\ full) = \frac{(x_i - C_r)^T V_r W_r V_r^T (x_i - C_r)}{\sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n}\left[(x_k - C_r)^T V_r W_r V_r^T (x_l - C_r)\right]^2}}, \qquad i = 1, \ldots, n, \tag{3.2}$$

where W_r is a diagonal matrix with jth diagonal element,

$$w_{rj} = \frac{1}{\max\{\lambda_{rj}, \lambda_{rs}\}}, \qquad j = 1, \ldots, p$$

and λ_{rs} is the smallest non-zero eigenvalue of S_r.

Step 2: Increase Size of Basic Subset

1. Rearrange the observations in ascending order according to equation (3.1) or (3.2) depending on whether S_r is of full rank or not.

2. Let r be the number of observations in the current basic subset.

   Divide the observations into two subsets:

   - a basic subset containing the first $r+1$ observations, and
   - a non-basic subset containing the remaining $n-r-1$ observations.

   That is, the size of the basic subset is increased by one and that of the non-basic subset is decreased by one.

3. Go to Step 3.

Step 3: Stopping Criterion

Compute $B_i(r)$ and b which will be defined later according to two different methods. If the smallest $B_i(r)$ is greater than $2b$, that is, if

$$\text{Min}B_i(r) > 2b\ , \qquad \forall i \in \text{non-basic subset}, \tag{3.3}$$

the procedure is stopped, else go to Steps 1 and 2. Two methods are considered in computing the values of $B_i(r)$ and b:

Method I: Basic Subset

All observations in the basic subset are considered in calculating the denominator and the numerator of b and the denominator of $B_i(r)$. Similar to Step 1, two cases, depending on whether the basic subset is of full rank or not, are considered.

Case 1: Basic Subset of Full Rank

If the sample covariance matrix S_r of the basic subset is of full rank, compute

$$b = \frac{\sum_{j=1}^{r}(x_j - C_r)^T S_r^{-1}(x_j - C_r)}{r\sqrt{\sum_{k=1}^{r}\sum_{l=1}^{r}\left[(x_k - C_r)^T S_r^{-1}(x_l - C_r)\right]^2}}, \qquad \forall x_j, x_k, x_l \in \text{basic subset}$$

and

$$B_i(r) = \frac{(x_i - C_r)^T S_r^{-1}(x_i - C_r)}{\sqrt{\sum_{k=1}^{r}\sum_{l=1}^{r}\left[(x_k - C_r)^T S_r^{-1}(x_l - C_r)\right]^2}}, \quad \begin{array}{l} \forall x_i \in \text{non-basic subset}, \\ \forall x_k, x_l \in \text{basic subset}, \\ i = 1, \ldots, n-r, \end{array} \tag{3.4}$$

where r is the number of observations in the final basic subset, C_r is the mean of the basic subset with r observations and S_r is the sample covariance matrix of the basic subset with r observations.

Case 2: Basic Subset Not of Full Rank

1. If S_r is not of full rank, compute the eigenvalues of S_r, $\lambda_{r1} \geq \ldots \geq \lambda_{rp} = 0$, and the corresponding normalized eigenvectors matrix V_r.

2. Compute

$$b = \frac{\sum_{j=1}^{r}(x_j - C_r)^T V_r W_r V_r^T (x_j - C_r)}{r\sqrt{\sum_{k=1}^{r}\sum_{l=1}^{r}\left[(x_k - C_r)^T V_r W_r V_r^T (x_l - C_r)\right]^2}}, \qquad \forall x_j, x_k, x_l \in \text{basic subset}$$

and

$$B_i(r) = \frac{(x_i - C_r)^T V_r W_r V_r^T (x_i - C_r)}{\sqrt{\sum_{k=1}^{r}\sum_{l=1}^{r}\left[(x_k - C_r)^T V_r W_r V_r^T (x_l - C_r)\right]^2}}, \quad \begin{array}{l} \forall x_i \in \text{non-basic subset}, \\ \forall x_k, x_l \in \text{basic subset}, \\ i = 1, \ldots, n-r, \end{array} \tag{3.5}$$

where W_r is a diagonal matrix with jth diagonal element,

$$w_{rj} = \frac{1}{\max\{\lambda_{rj}, \lambda_{rs}\}}, \qquad j = 1, \ldots, p$$

and λ_{rs} is the smallest non-zero eigenvalue of S_r.

Method II: All Observations

The difference between Method I and Method II is that all observations in both basic subset and non-basic subset are considered in calculating the denominator and the numerator of b and the denominator of $B_i(r)$ in this method. The equations of b and $B_i(r)$ are the same as those in Method I, except

$\forall x_k, x_l \in$ basic subset	is changed to	$\forall x_k, x_l \in$ set of all observations
$\forall x_j, x_k, x_l \in$ basic subset	is changed to	$\forall x_j, x_k, x_l \in$ set of all observations.

Similar to Method I, two cases are also considered for this method.

Case 1: Basic Subset of Full Rank

If S_r is of full rank, compute

$$b = \frac{\sum_{j=1}^{n}(x_j - C_r)^T S_r^{-1}(x_j - C_r)}{n\sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n}\left[(x_k - C_r)^T S_r^{-1}(x_l - C_r)\right]^2}}, \qquad \forall x_j, x_k, x_l \in \text{set of all observations}$$

and

$$B_i(r) = \frac{(x_i - C_r)^T S_r^{-1}(x_i - C_r)}{\sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n}\left[(x_k - C_r)^T S_r^{-1}(x_l - C_r)\right]^2}}, \quad \begin{array}{l} \forall x_i \in \text{non-basic subset}, \\ \forall x_k, x_l \in \text{set of all observations}, \\ i = 1, \ldots, n-r. \end{array} \tag{3.6}$$

Case 2: Basic Subset Not of Full Rank

If S_r is not of full rank, compute

$$b = \frac{\sum_{j=1}^{n}(x_j - C_r)^T V_r W_r V_r^T (x_j - C_r)}{n\sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n}\left[(x_k - C_r)^T V_r W_r V_r^T (x_l - C_r)\right]^2}}, \qquad \forall x_j, x_k, x_l \in \text{set of all observations}$$

and

$$B_i(r) = \frac{(x_i - C_r)^T V_r W_r V_r^T (x_i - C_r)}{\sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n}\left[(x_k - C_r)^T V_r W_r V_r^T (x_l - C_r)\right]^2}}, \quad \begin{array}{l} \forall x_i \in \text{non-basic subset}, \\ \forall x_k, x_l \in \text{set of all observations}, \\ i = 1, \ldots, n-r. \end{array} \tag{3.7}$$

The forms of $B_i(r)$ for the two methods are similar to the Mahalanobis distance but are standardized by constants which are the square root parts of the denominators of $B_i(r)$ (see equations (3.4), (3.5), (3.6), (3.7)) to make the values of $B_i(r)$ between zero and one. The observations from the basic subset and observations from the set of all observations, which includes both the basic subset and the non-basic subset, are used to calculate the constants or the square root parts of $B_i(r)$ in Method I and Method II respectively. No observations from the non-basic subset are used to calculate the square root parts of $B_i(r)$ because the observations are considered as outliers of the whole data set and are not suitable for the standardization.

The observations in the final non-basic subset are considered as outliers if (3.3) is satisfied. That is, when the minimum of the distances between observations in the non-basic subset and the mean of the basic subset relative to its sample covariance matrix is greater than the benchmark $2b$, the observations in the non-basic subset are far away from the majority of data and are declared as outliers.

3.2 Examples

Two C language programs with double precision are written to implement the new procedure proposed in Section 3.1. One program is written for Method I and the other is written for Method II. Several reported data sets are applied to the new procedure in this section. The new procedure is applied to the following data sets:

1. Hawkins, Bradu and Kass data set from Hawkins, Bradu and Kass (1984)

2. Brain and body weight data set from Rousseeuw & Leroy (1987, p.58)

3. Stack loss data set from Brownlee (1965)

4. An artificial data set from Poon, Lew & Poon (2000)

5. An artificial multivariate normal data set with sample size $n = 200$, dimension $p = 40$, fraction of contamination $\varepsilon = 0.05$, constant defining the amount of shift for location outliers $d = 2$. This data set is constructed by the method of generating simulation data set proposed by Rocke & Woodruff (1996). The method of constructing this data set is described below and the details of constructing simulation data set are described in the next section.

The artificial multivariate normal data set is generated by using a C language program and the steps of constructing the data set are as follows:

1. Put

sample size	$n = 200$,
dimension	$p = 40$,
fraction of contamination	$\varepsilon = 0.05$, and
constant defining the amount of shift for close outliers	$d = 2$

where sample size n is the number of observations of the data set, dimension p is the number of variables in the data set, fraction of contamination ε is the proportion of outliers in the data set and constant d defining the amount of shift for close outliers is a constant for constructing close outliers. The details of the above terms will be explained in the next section.

2. Draw $n(1-\varepsilon)$ good observations from a multivariate normal distribution $N(0, I)$ with zero mean and covariance matrix I where I is an identity matrix, and draw $n\varepsilon$ bad observations from a multivariate normal distribution $N(dQ_p^*, I)$ with mean dQ_p^* and covariance matrix I where $Q_p^* = \sqrt{\chi^2_{p,0.999}/p}$ and $\chi^2_{p,0.999}$ is the 0.999 probability point of a chi-squared distribution with p degrees of freedom.

The first three reported data sets are chosen as they are typical examples of identifying multivariate outliers in the literature. The fourth data set is chosen as it is an artificial data set for testing the outlier measure B_i as shown in Poon, Lew & Poon (2000). The data set is used to assess the performance of the new procedure by using the outlier measure. For the fifth data set, the sample size

n, the dimension p, the number of outliers $n\varepsilon$ and the constant d defining the amount of shift are the same as those of the high dimensional artificial data set constructed by Poon, Lew & Poon (2000). The data set is used to study whether the proposed procedure is effective in dealing with high dimensional data. The above data sets are used to compare the performance of the new procedure to other algorithms of detecting multivariate outliers in the literature.

The results are presented by using index plots to show the outlying observations. Index plot is a plot of the outlier measures B_i's of observations versus the observation numbers. "o" on the index plot indicates that the observation is an outlier. Some of the outliers found by Method II are also labelled by the corresponding observation numbers.

Example 1: Hawkins, Bradu and Kass data set

The first data set is constructed by Hawkins, Bradu and Kass (1984) with sample size $n = 75$ and dimension $p = 3$. The first 14 observations are detected as outliers by Rousseeuw & van Zomeren (1990), Hadi (1992), Atkinson & Mulira (1993) and Poon, Lew & Poon (2000). The first 10 observations form a group of outliers while the other four observations form another group. Figure 3.1 shows the index plots of B_i's for Method I and Method II. For Method I, all the first 14 observations are declared as outliers with observations 11 to 14 as the most extreme ones as shown in Figure 3.1(a). But other observations, except four observations (67, 59, 29, 50. All the observations written in this form are ordered

in descending order according to B_i), are also treated as outliers. Figure 3.1(b) shows that two groups of outliers appear for Method II. The first group of outliers includes observations 1 to 10 and observations 11 to 14 form the second group. The result of using Method II is what we expected. However, Method I is not appropriate in identifying outliers in this example as nearly all the observations are declared as outliers.

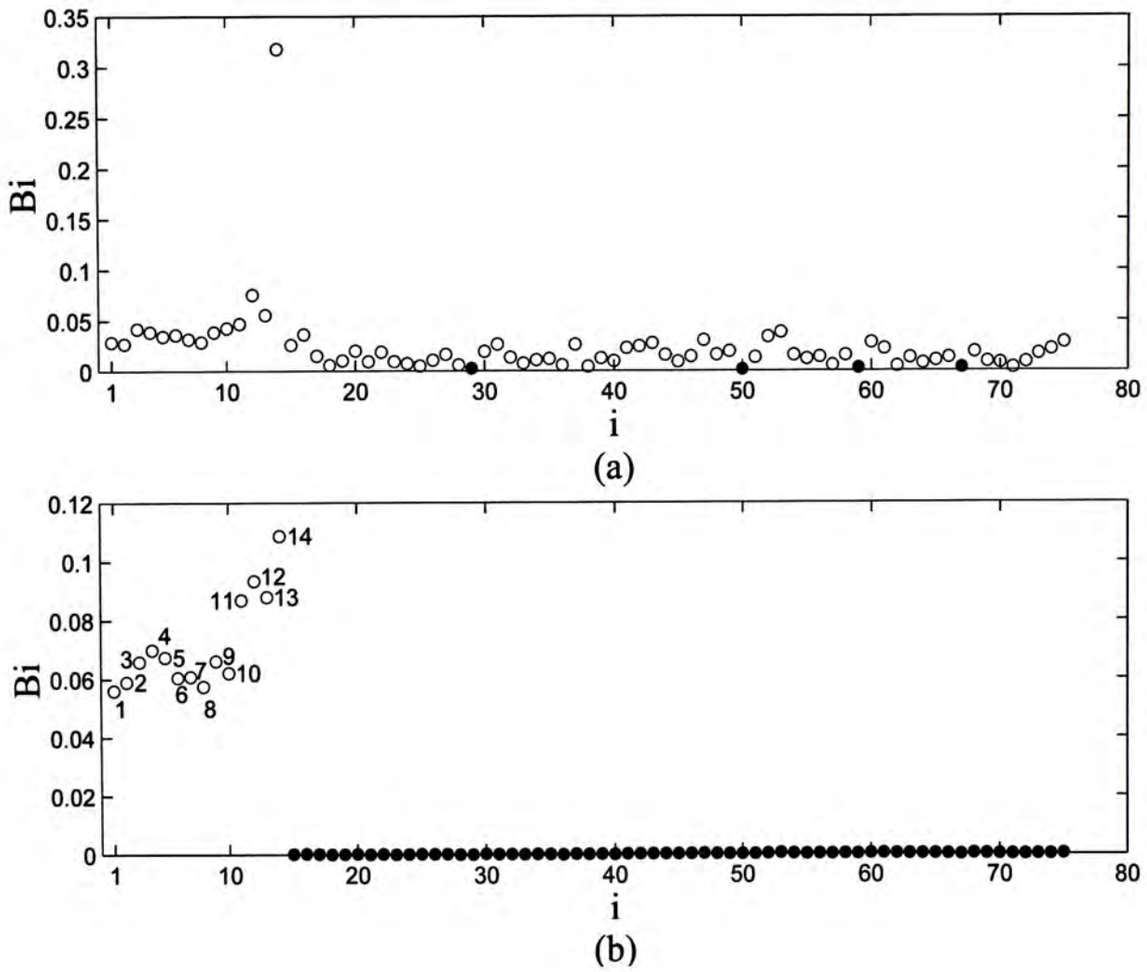

Figure 3.1: Index plot for the Hawkins, Bradu and Kass data using the sample covariance matrix and (a) Method I, (b) Method II

Example 2: Brain and body weight data set

Brain and body weight data set is taken from Rousseeuw & Leroy (1987, p.58). The data set shows the relationship between the brain weight and the body weight of 28 different animals. The data were taken logarithms to the base 10 so that the relationship between the brain weight and the body weight can be clearly shown as mentioned in Rousseeuw & Leroy (1987). Hadi (1992), Rousseeuw &

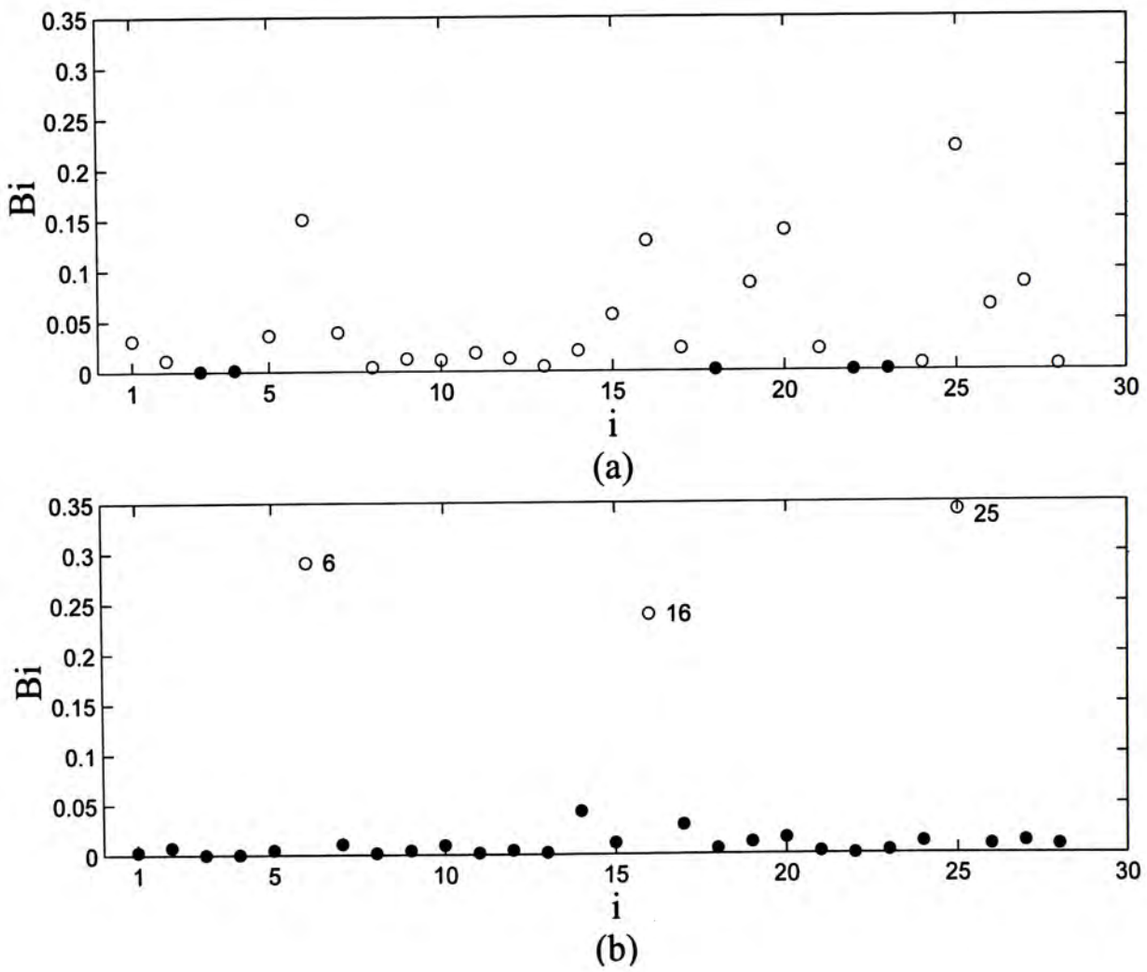

Figure 3.2: Index plot for the brain and body weight data using the sample covariance matrix and (a) Method I, (b) Method II

van Zomeren (1990) and Atkinson & Mulira (1993) declare that observations of three dinosaurs 25, 6 and 16 are outliers. Observations 14 and 17, which are observations of human beings and rhesus monkeys respectively, are considered as marginal cases. The index plots of B_i's are shown in Figure 3.2. Observations 25, 6, 16 are declared as outliers for both methods. However, 23 observations including the three outliers mentioned are shown to be outliers by Method I in Figure 3.2(a). The five largest values of B_i's for Method I and Method II are in the order 25, 6, 20, 16, 19 (Figure 3.2(a)) and 25, 6, 16, 14, 17 (Figure 3.2(b)) respectively. Observations 14 and 17 are not included for Method I but for Method II. Method I again identifies too many observations as outliers than it should be.

Example 3: Stack loss data set

Brownlee's (1965, p.454) stack loss data is used in this example. The data set is obtained from operation of a plant for the oxidation of ammonia to nitric acid. The three explanatory variables $p = 3$ with 21 observations are used in this example. Hadi (1992) detects four outliers 2, 1, 3 and 21 in this order but Atkinson (1986) and Atkinson (1994) declare observations 1, 3, 4 and 21 as outliers. The outliers found by Method I or Method II are not the same as the outliers mentioned above. The index plot for Method I is given in Figure 3.3(a). Observations 21 and 4 are two of the 15 outliers but they are not among the first four most extreme ones. Observations 1 to 3 are treated as good observations with smaller values of B_i than other observations. Figure 3.3(b) indicates that

observation 17 is the only outlier by using Method II and observations 21, 2 and 1 have large values of B_i. Most of the observations that are far away from the bulk of data as mentioned in the literature are detected by using Method II but more than half of the observations are identified as outliers by using Method I.

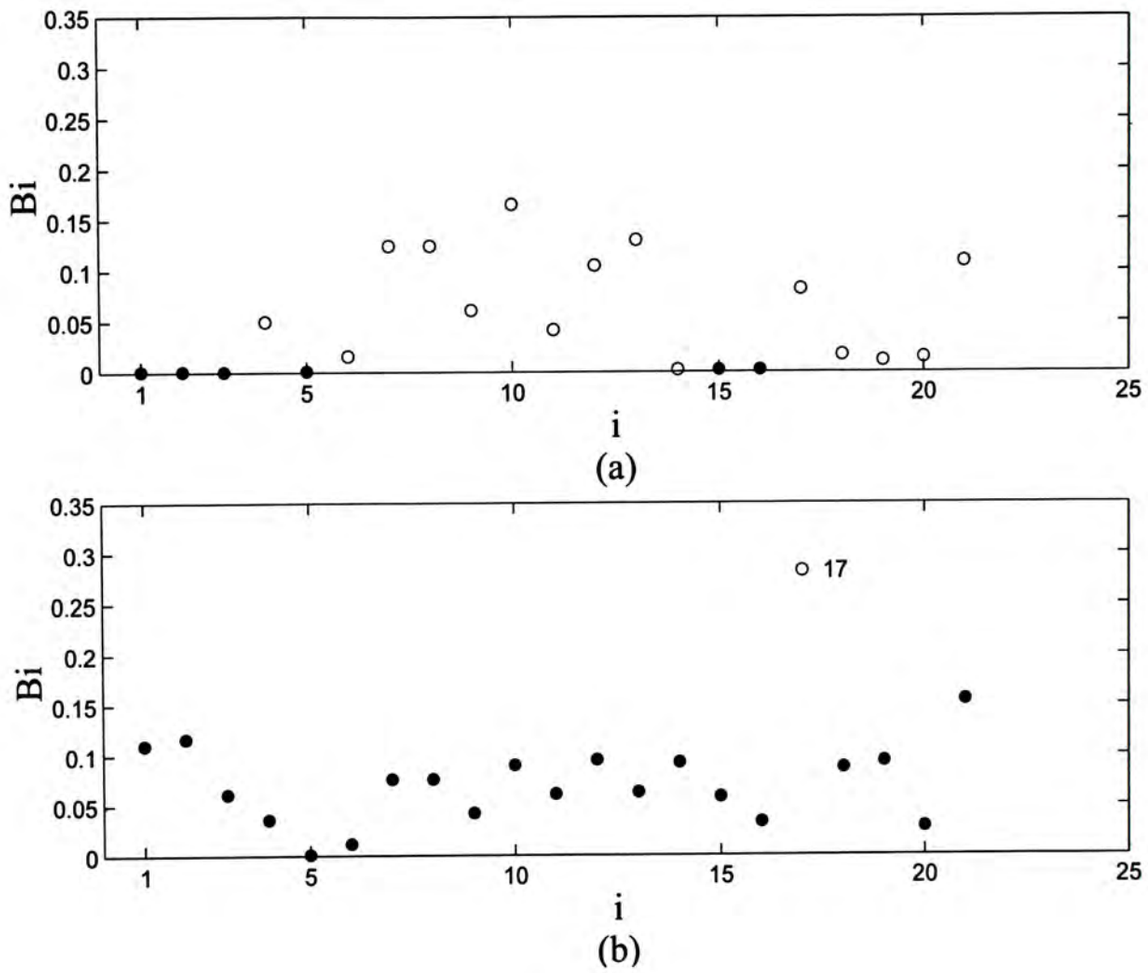

Figure 3.3: Index plot for the stack loss data set using the sample covariance matrix and (a) Method I, (b) Method II

Example 4: An artificial data set

This artificial data set is constructed by Poon, Lew & Poon (2000). Observations 9 and 10 are constructed as outlying observations. For Method I, seven observations 9, 10, 1, 8, 2, 7, 3 which are ordered in descending order according to B_i are identified as outliers in Figure 3.4(a). For Method II, only observation 9 is detected as an outlier in Figure 3.4(b). Observations 9 and 10 have the largest

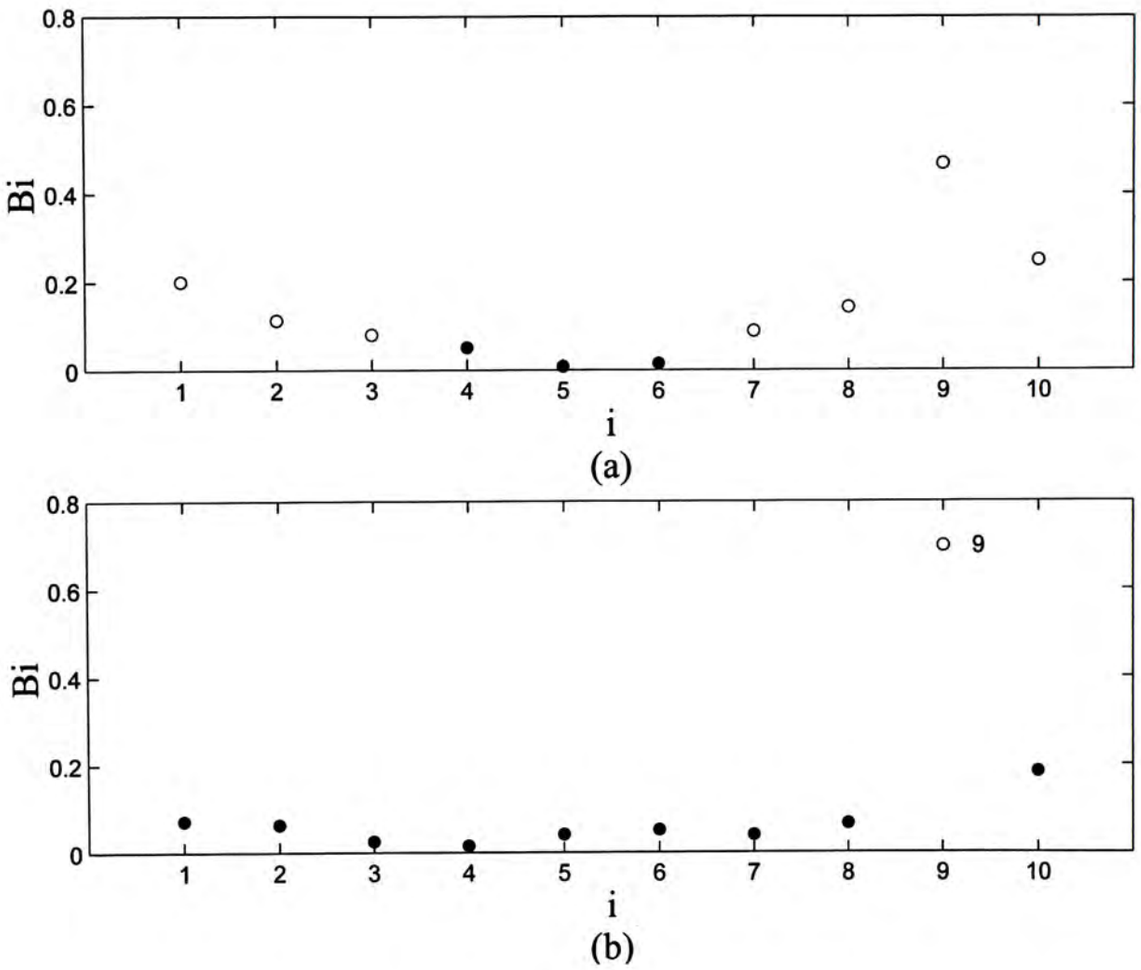

Figure 3.4: Index plot for the artificial data using the sample covariance matrix and (a) Method I, (b) Method II

value and the second largest value of B_i respectively for both methods. Method II gives reasonable results but not Method I as Method I detects more than half of the observations as outliers.

Example 5: An artificial multivariate normal data set

The construction of this artificial multivariate normal data set is described at the beginning of this section. The same kind of data set is also constructed by

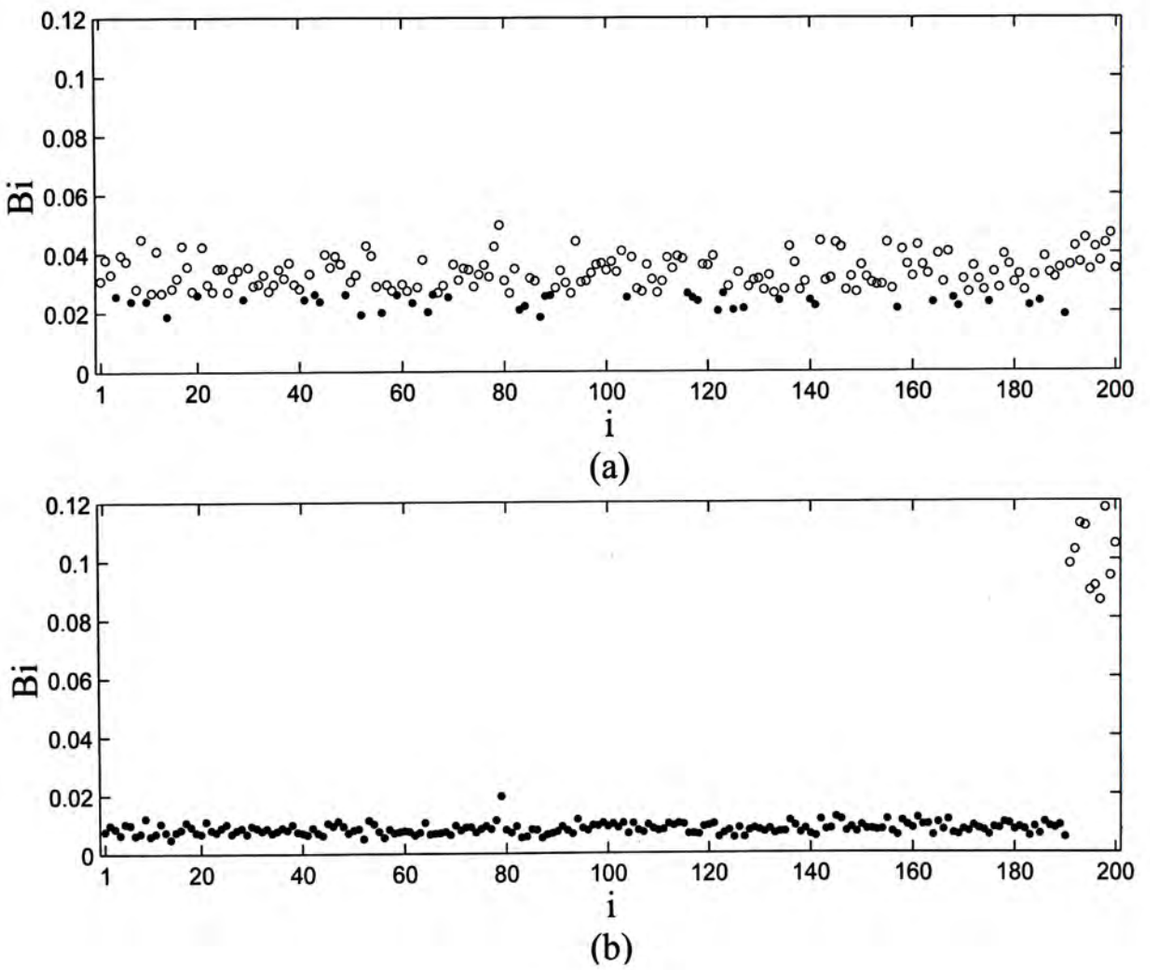

Figure 3.5: Index plot for the artificial multivariate normal data using the sample covariance matrix and (a) Method I, (b) Method II

Poon, Lew & Poon (2000). The last 10 observations are constructed as outliers. They use the data set to show the effectiveness of their outlier measure when the value of sparseness n/p is small. The ratio n/p is a measure of the number of observations relative to the size of the dimension. When dimension increases, the number of observations must be increased to make the ratio n/p large. Outliers can be detected more easily when the ratio is large. Figure 3.5 gives the index plots of B_i's. For Method I, 159 observations are declared as outliers in Figure 3.5(a). Within the 159 observations, nine of the 10 constructed outliers are detected as outliers. All the last 10 observations are identified as outliers for Method II as shown in Figure 3.5(b). This example shows that the new procedure is successful in detecting outliers from high dimensional data by Method II but not Method I because too many observations are declared as outliers by using Method I.

From the above examples, it is clear that Method I of the new procedure should be rejected because it usually identifies more than half of the observations as outliers, leading to a very high misclassification rate which is defined as the ratio of the number of good points detected as outliers to the total number of good points generated. Since only the observations in the basic subset are used to calculate the benchmark $2b$ for Method I, the value of b is usually very small and the stopping criterion (3.3) is likely to be met earlier than expected and hence the number of observations in the final basic subset is small. This leads to the identification of more than half of the observations as outliers. This situation will

be explained further in the discussion.

3.3 Simulation Study

In this section, simulation study is performed to assess the performance of Method II of the proposed procedure. Method I is not considered here because its misclassification rate is very high. We believe that the performance of the proposed procedure is affected by a number of factors: the nature of multivariate outliers, amount of shift, fraction of contamination, dimension and sample size. Some terms related to the simulation study and the factors mentioned above are described first in this section. The procedure and the results of the simulation study are given in the latter parts of this section. All the distributions in the simulation study are multivariate normal distributions.

3.3.1 Terms and Factors

Some terms given by Rocke & Woodruff (1996) which are related to the simulation study are summarized as follows:

- Good data points — are the majority of the data coming from a well-behaved population, for example, multivariate standard normal population.

- Bad data points — are the remainders of the data not coming from the well-behaved population. They may be drawn from a displaced population. In other words, they come from a population with the same covariance matrix as that of good data but with a shifted mean.

- Shape — from definition 3 of Rocke & Woodruff (1996),

  let X be a $n \times p$ matrix representing a sample of n points in $\mathbf{R}^p$ where n is the sample size and p is the dimension. Let $S = n^{-1}(X - \overline{X})^T(X - \overline{X})$ be the sample covariance matrix. The shape of X is $S/|S|^{1/p}$.

Factors

Rocke & Woodruff (1996) have demonstrated that several factors would affect the performance of location outlier identification procedure. Those factors are likely to affect the performance of the proposed procedure and will be considered in our simulation studies. They are described as in the following:

Nature of Multivariate Outliers

Consider a sample of n points coming from a multivariate normal population in $\mathbf{R}^p$. Let the good data points come from a multivariate normal population $N(\mu_0, \Sigma_0)$ with mean μ_0 and covariance matrix Σ_0. Let the bad data points come from a multivariate normal population $N(\mu_0 + \mu, \Omega) = N(\mu_0 + \mu, \lambda\Sigma_0)$ with mean $\mu_0 + \mu$ and covariance matrix $\Omega = \lambda\Sigma_0$ where μ is a constant vector and λ is a constant. If $\mu_0 = 0$, μ is the amount of shift to form a displaced population. If $\lambda = 1$, pure shift outliers generated from $N(\mu_0 + \mu, \Sigma_0)$ form the bad data.

By Theorem 1 of Rocke & Woodruff (1996), outliers with the same shape as the good data are the hardest to find. That is, pure shift outliers are the hardest to find. If pure shift outliers can be identified, other kinds of outliers will also

be detected. Therefore, pure shift outliers coming from $N(\mu_0 + \mu, \Sigma_0)$ and good data coming from $N(\mu_0, \Sigma_0)$ are considered in the simulation study. The amount of shift μ and the number of bad data points or fraction of contamination are explained in the following.

Amount of Shift

The amount of shift μ is measured in terms of Q_p. Consider a sphere ,which contains most of the good data and centers at mean μ_0, with radius $Q_p = \sqrt{\chi^2_{p,0.999}}$ where $\chi^2_{p,0.999}$ is the 0.999 probability point of a chi-squared distribution with p degrees of freedom. The amount of shift μ equals dQ_p^* where $Q_p^* = \sqrt{\chi^2_{p,0.999}/p}$ and d is a constant defining the amount of shift. When d is small, outliers are considered as close outliers and when d is large, outliers are far away from the bulk of data. Far outliers are easier to be detected than close outliers.

Fraction of Contamination

Fraction of contamination ε is the proportion of outliers in the whole data set. Rocke & Woodruff (1996) show that if fraction of contamination ε increases, the success rate of detecting outliers will decrease. If fraction of contamination is greater than $\frac{1}{p+1}$ where p is the dimension of the data, Rocke & Woodruff (1996) reported that most methods of identifying outliers from the literature will break down. Rocke & Woodruff (1996) also point out that the number of outliers should be less than half of the observations in a data set and more than half of the data should come from a well-behaved population. On the other hand, according to

Lopuhaä & Rousseeuw (1991), the number of good data points should be at least $h = [(n + p + 1)/2]$, the integer part of $(n + p + 1)/2$. Hence, fraction of contamination ε should be less than $(n - h)/n$ which is the proportion of bad data points. Note that h is the integer part of $(n + p + 1)/2$ and satisfies

$$\#\{i : (x_i - \mathbf{t_n})^T \mathbf{C_n}^{-1} (x_i - \mathbf{t_n}) \leq c^2\} \geq \left[\frac{n + p + 1}{2}\right],$$

where $\mathbf{x} = \{x_1, \ldots, x_n\}$ is a collection of $n > p + 1$ points in $\mathbf{R}^p$, $\mathbf{t_n}$ is a location estimate based on $\mathbf{x}$, $\mathbf{C_n}$ is a covariance estimate, and c is a constant and equals $\chi^2_{p,0.50}$ if a normal sample is considered.

$\mathbf{t_n}$ and $\mathbf{C_n}$ determine the center and the covariance structure of the minimum volume ellipsoid covering at least $[(n + p + 1)/2]$ points so that the fraction of outliers will not spoil the estimate of location and covariance matrix as described in Lopuhaä & Rousseeuw (1991).

Dimension

When dimension p increases, the proportion of outliers that can be handled decreases even with large sample size. For example, cases with dimension 40 or higher will still be manageable if the amount of contamination is small enough (20–25%) from Rocke & Woodruff (1996).

Sample Size

Sample size n should increase when dimension p increases. For fixed d and p, and all levels of contamination, Rocke & Woodruff (1996) show that the success rate of detecting outliers increases with sample size.

3.3.2 Procedure

Consider a sample of size n with dimension p, coming from a multivariate normal population. Let the good data come from a multivariate standard normal population $N(\mu_0, \Sigma_0) = N(0, I)$ with mean $\mu_0 = 0$ and covariance matrix $\Sigma_0 = I$ where I is an identity matrix. Let the bad data, which form a fraction ε of the whole data, come from $N(\mu_0 + \mu, \Sigma_0) = N(\mu, I)$ when pure shift outliers are used.

The simulation study is conducted according to the following steps:

1. Choose n and p (≤ 40) such that n grows with p and $n > p$. Here we choose

$$\begin{aligned} n &= 100, 200, 400, 800 \qquad \text{and} \\ p &= 5, 10, 20. \end{aligned}$$

2. Choose

$$\varepsilon = 0.2, 0.3, 0.4.$$

The values of the above ε's are greater than the value $\frac{1}{p+1}$ as described in the criterion $\varepsilon < \frac{1}{p+1}$ and less than the value $\frac{n-h}{n}$ as described in the criterion

$\varepsilon < \frac{n-h}{n}$ aforementioned. The purpose of choosing these values is to examine the performance of the proposed procedure at various contamination rates.

3. Report

   (a) total number of outliers generated $= n\varepsilon$, and

   (b) total number of good points generated $= n(1-\varepsilon)$.

4. (a) Use $d = 4$ to get far outliers.

   (b) Use $d = 2$ to get close outliers.

5. Draw $n(1-\varepsilon)$ good data points from $N(\mu_0, \Sigma_0) = N(0, I)$ and $n\varepsilon$ bad data points from $N(\mu_0+\mu, \lambda\Sigma_0) = N(dQ_p^*, I)$, where $\lambda = 1$ for pure shift outliers, $Q_p^* = \sqrt{\chi^2_{p,0.999}/p}$ and

$$\begin{aligned} \text{mean of bad data} &= \mu_0 + \mu \\ &= \mu_0 + dQ_p^* \\ &= 0 + dQ_p^* \\ &= dQ_p^*. \end{aligned}$$

6. Apply Method II of the proposed procedure to the above dataset.

7. Report

   (a) number of data points detected as outliers,

   (b) number of good points detected as outliers, and

(c) number of outliers detected without misclassification

= number of data points detected as outliers − number of good points detected as outliers.

8. Calculate the misclassification rate MR_i and the success rate without misclassification SR_i for each trial $i, i = 1, \ldots, 50$, where

$$MR_i = \frac{\text{number of good points detected as outliers in trial } i}{\text{total number of good points generated}}$$

and

$$SR_i = \frac{\text{number of outliers detected without misclassification in trial } i}{\text{total number of outliers generated}}.$$

9. Repeat steps 5–8 for 50 trials for fixed n, p, ε and d.

10. Calculate the average misclassification rate AMR and the average success rate without misclassification ASR, where

$$AMR = \frac{\sum_{i=1}^{50} MR_i}{50}$$

and

$$ASR = \frac{\sum_{i=1}^{50} SR_i}{50}.$$

3.3.3 Results

The results by using Method II of the proposed procedure are presented in four tables in the Appendix, Table 1 under column (a) to Table 4 under column

(a). Tables 1 to 4 carry similar information but they are reorganized to facilitate comparison on the effect of a given factor. For each table, three factors from n, p, μ (or d) and ε are fixed and the effect of the remaining factor can be seen. The average misclassification rates and the average success rates without misclassification are given in the tables.

The procedure is considered to be reliable and satisfactory if the value of the average success rate without misclassification ASR is large and the value of the average misclassification rate AMR is small. The ASR's are highlighted in each row box for both columns (a) and (b) if more than half of the ASR's in each row box are greater than 0.8 for either column (a) or (b). We will pay more attention to the highlighted ASR's and the corresponding AMR's in the tables. Other cases are not emphasized as the ASR's are too small to be reliable.

Effect of Dimension

Column (a) of Table 1 shows the effect of dimension p with n, d and ε fixed. The ASR's are highlighted when $\varepsilon = 0.2$ or $\varepsilon = 0.3$ and $d = 4$. Most of the highlighted ASR's decrease when p increases which is what we expected, except those underlined ASR's showing a slightly increasing trend. However, the underlined ASR's are very large (> 0.99) and the increase in the trend is small (≤ 0.003). We expected that all the underlined values are large and the increase in the trend is due to some random effects in generating the simulated data sets. Column (a) of Table 1 also shows that the AMR's increase with dimension p for

the highlighted cases and the AMR's for the underlined cases are zero correcting to the 6 decimal places. Therefore, the slightly increase in the trend is negligible and all the underlined ASR's are considered to be large and reliable.

Effect of Sample Size

Column (a) of Table 2 gives the effect of sample size n with d, p and ε fixed. The highlighted ASR's are those cases with $\varepsilon = 0.2$ or $\varepsilon = 0.3$, $d = 4$ and $p = 5$ or 10. Column (a) of Table 2 shows that the AMR decreases with increasing sample size for the highlighted cases. We expected that the ASR increases with sample size. All the highlighted ASR's show an increasing trend when the sample size increases, except the underlined ASR's. However, all the underlined ASR's for $p = 5, d = 2$ and $\varepsilon = 0.2$ and those for $p = 5, d = 4$ and $\varepsilon = 0.3$ are greater than 0.99 and 0.98 respectively. Both are considered to be large. Besides, the AMR's for the underlined cases are small (< 0.0015). We expected that the reason why the underlined ASR's do not follow an increasing trend is due to some random effects in generating the data sets. Thus, the underlined ASR's are considered to be large and the results of all the highlighted cases provide the evidence that the ASR increases with sample size.

Effect of Amount of Shift

The effect of amount of shift μ shown by the constant d with n, p and ε fixed is shown in column (a) of Table 3. We expected that the ASR's for far outliers with $d = 4$ are larger than those for close outliers with $d = 2$. Most of the ASR's with

ε fixed at 0.2 are highlighted and match with what we expected. The ASR's for far outliers and those for close outliers are nearly the same when ε is fixed to be 0.2. The highlighted ASR's for both $d = 2$ and $d = 4$ are greater than 0.9 and the corresponding AMR's are small. However, the underlined ASR for far outliers is slightly less than that for close outliers. As the ASR's for the underlined cases are large (> 0.9) and the corresponding AMR's are small (< 0.0015), the slightly decrease in ASR's for far outliers is negligible. Therefore, the effect of amount of shift on the proposed procedure matches with our expectation.

Effect of Fraction of Contamination

Column (a) of Table 4 shows the effect of fraction of contamination ε when n, d and p are fixed. We expected that the ASR decreases when the fraction of contamination increases. Column (a) of Table 4 shows that the highlighted ASR's with $d = 4$ and $p = 5$ or 10, and $n = 800, p = 20$ and $d = 4$ decrease with increasing AMR's when ε increases with other factors fixed. The results of the effect of fraction of contamination match with what we expected.

From the above outcomes, most of the satisfactory results appear when the fraction of contamination ε equals 0.2. When ε increases, the average success rate without misclassification decreases drastically and the average misclassification rate increases. This indicates that the fraction of contamination is a significant factor in affecting the values of ASR's and AMR's.

Chapter 4

Robust Version of the New Procedure

The average success rates without misclassification are not satisfactory when the fraction of contamination ε increases in the simulation study. Also, sample mean and sample covariance matrix are highly affected by outlying observations. So, a refinement is made on the proposed procedure in this chapter. Robust estimate of the location C_M and robust estimate of the sample covariance matrix S_M are used to replace the sample mean $\hat{\mu}$ and the sample covariance matrix S in the outlier measure B_i respectively to eliminate the effect caused by outlying observations. The robust version of the procedure is shown in Section 4.1. Similar to Chapter 3, some reported data sets and simulation study are used to show the effectiveness of the revised procedure in Sections 4.2 and 4.3 respectively.

4.1 Procedure

The revised procedure using the robust estimates is similar to the procedure using the sample covariance matrix. The difference is that a co-ordinatewise median vector C_M is used to estimate the population mean vector in B_i (see equations (4.2) and (4.3)). The procedure using the robust estimates is as follows:

Step 0: Initial Ordering

1. Compute C_M and S_M, where C_M is the co-ordinatewise median vector of

$$\mathbf{X_{n\times p}} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}$$

i.e.

$$C_M = \begin{pmatrix} C_{M1} \\ \vdots \\ C_{Mp} \end{pmatrix}, \tag{4.1}$$

where C_{Mj} is the median of the elements in the vector

$$\begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix}, \qquad j = 1, \dots, p$$

and

$$S_M = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - C_M)(x_i - C_M)^T. \tag{4.2}$$

2. Rearrange the n observations in ascending order according to B_i, where

$$B_i = \frac{(x_i - C_M)^T S_M^{-1}(x_i - C_M)}{\sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n}\left[(x_k - C_M)^T S_M^{-1}(x_l - C_M)\right]^2}}, \qquad i = 1, \ldots, n. \tag{4.3}$$

3. Divide the observations into two initial subsets according to the ascending order of the observations:

   - a basic subset containing the first $p+1$ observations, and
   - a non-basic subset containing the last $n-p-1$ observations.

4. Go to Step 3 and check the stopping criterion before go to Steps 1 and 2.

Step 1: Use Basic Subset

In this step, the co-ordinatewise median vector C_{Mr} and the sample robust covariance matrix S_{Mr} of the basic subset are used to replace C_M and S_M respectively in calculating B_i (equation (4.3)) in Step 0. Two cases, depending on whether the sample robust covariance matrix of the basic subset is of full rank or not, are considered in this step.

Case 1: Basic Subset of Full Rank

If the sample robust covariance matrix S_{Mr} of the basic subset is of full rank, compute

$$B_i(full) = \frac{(x_i - C_{Mr})^T S_{Mr}^{-1}(x_i - C_{Mr})}{\sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n}\left[(x_k - C_{Mr})^T S_{Mr}^{-1}(x_l - C_{Mr})\right]^2}}, \qquad i = 1, \ldots, n, \quad (4.4)$$

where C_{Mr} and S_{Mr} are the co-ordinatewise median vector and the robust covariance matrix of the basic subset.

Case 2: Basic Subset Not of Full Rank

1. If S_{Mr} is not of full rank, compute the eigenvalues of S_{Mr}, $\lambda_{Mr1} \geq \ldots \geq \lambda_{Mrp} = 0$, and the corresponding normalized eigenvectors matrix V_{Mr}.

2. Compute

$$B_i(not\ full) = \frac{(x_i - C_{Mr})^T V_{Mr} W_{Mr} V_{Mr}^T (x_i - C_{Mr})}{\sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n}\left[(x_k - C_{Mr})^T V_{Mr} W_{Mr} V_{Mr}^T (x_l - C_{Mr})\right]^2}}, \quad i = 1, \ldots, n, \quad (4.5)$$

where W_{Mr} is a diagonal matrix with jth diagonal element,

$$w_{Mrj} = \frac{1}{\max\{\lambda_{Mrj}, \lambda_{Mrs}\}}, \qquad j = 1, \ldots, p$$

and λ_{Mrs} is the smallest non-zero eigenvalue of S_{Mr}.

Step 2: Increase Size of Basic Subset

1. Rearrange the observations in ascending order according to equation (4.4) or (4.5) depending on whether S_{Mr} is of full rank or not.

2. Let r be the number of observations in the current basic subset.

   Divide the observations into two subsets:

   - a basic subset containing the first $r+1$ observations, and
   - another subset containing the remaining $n-r-1$ observations.

3. Go to Step 3.

Step 3: Stopping Criterion

If the smallest $B_i(r)$ is greater than $2b$, i.e.

$$\text{Min}B_i(r) > 2b\ , \qquad \forall i \in \text{non-basic subset}, \tag{4.6}$$

stop the procedure, else go to Steps 1 and 2.

As Method I is rejected in Chapter 3, only Method II is used to find $B_i(r)$ and b as follows:

Method II: All Observations

Case 1: Basic Subset of Full Rank

If S_{Mr} is of full rank, compute

$$b = \frac{\sum_{j=1}^{n}(x_j - C_{Mr})^T(S_{Mr})^{-1}(x_j - C_{Mr})}{n\sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n}\left[(x_k - C_{Mr})^T(S_{Mr})^{-1}(x_l - C_{Mr})\right]^2}},$$

$\forall x_j, x_k, x_l \in$ set of all observations

and

$$B_i(r) = \frac{(x_i - C_{Mr})^T(S_{Mr})^{-1}(x_i - C_{Mr})}{\sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n}\left[(x_k - C_{Mr})^T(S_{Mr})^{-1}(x_l - C_{Mr})\right]^2}},$$

$\forall x_i \in$ non-basic subset,

$\forall x_k, x_l \in$ set of all observations,

$i = 1, \ldots, n - r,$

where r is the number of observations in the final basic subset, C_{Mr} is the median of the basic subset with r observations and S_{Mr} is the sample robust covariance matrix of the basic subset with r observations.

Case 2: Basic Subset Not of Full Rank

1. If S_{Mr} is not of full rank, compute the eigenvalues of S_{Mr}, $\lambda_{Mr1} \geq \ldots \geq \lambda_{Mrp} = 0$, and the corresponding normalized eigenvectors matrix V_{Mr}.

2. Compute

$$b = \frac{\sum_{j=1}^{n}(x_j - C_{Mr})^T V_{Mr} W_{Mr} V_{Mr}^T (x_j - C_{Mr})}{n\sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n}\left[(x_k - C_{Mr})^T V_{Mr} W_{Mr} V_{Mr}^T (x_l - C_{Mr})\right]^2}},$$

$\forall x_j, x_k, x_l \in$ set of all observations

and

$$B_i(r) = \frac{(x_i - C_{Mr})^T V_{Mr} W_{Mr} V_{Mr}^T (x_i - C_{Mr})}{\sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n}\left[(x_k - C_{Mr})^T V_{Mr} W_{Mr} V_{Mr}^T (x_l - C_{Mr})\right]^2}},$$

$\forall x_i \in$ non-basic subset,

$\forall x_k, x_l \in$ set of all observations,

$i = 1, \ldots, n - r,$

where W_{Mr} is a diagonal matrix with jth diagonal element,

$$w_{Mrj} = \frac{1}{\max\{\lambda_{Mrj}, \lambda_{Mrs}\}}, \qquad j = 1, \ldots, p$$

and λ_{Mrs} is the smallest non-zero eigenvalue of S_{Mr}.

The observations in the final non-basic subset are considered as outliers if (4.6) is satisfied.

4.2 Examples

The revised procedure using Method II is applied to the same reported data sets used in Chapter 3. The results are also presented by index plots. Here are the results:

Example 1: Hawkins, Bradu and Kass data set

Similar to the results found in Chapter 3, the index plot in Figure 4.1 shows two groups of outliers. The first 10 observations form a group of outliers and observations 11 to 14 form another group of outliers. The values of B_i in Figure 3.1(b) by using the sample covariance matrix are nearly the same as those values in Figure 4.1 by using the sample robust covariance matrix. Using Method II, both the proposed procedure using the sample covariance matrix and the revised procedure using the sample robust covariance matrix are effective in detecting

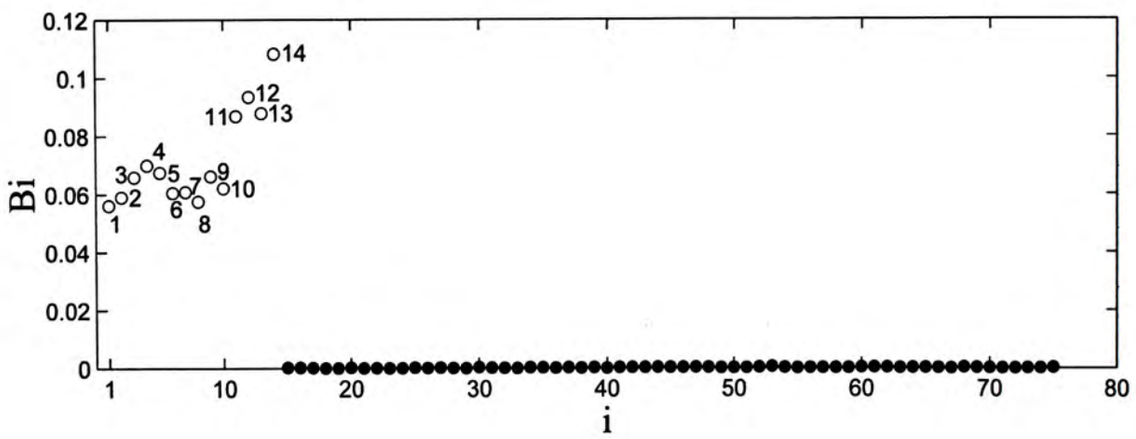

Figure 4.1: Index plot for the Hawkins, Bradu and Kass data using Method II and the robust estimate

the 14 outliers and the results of both procedures match the results in the literature.

Example 2: Brain and body weight data set

Observations 25, 6 and 16 are declared as outliers as shown in Figure 4.2. The outliers found are the same as those found in Chapter 3 as shown in Figure 3.2(b). Figures 3.2(b) and 4.2 indicate that the values of B_i obtained from the two proposed procedures in the last chapter and this chapter are similar. The two proposed procedures make no difference to the results. Also, the results are consistent with the results in the literature.

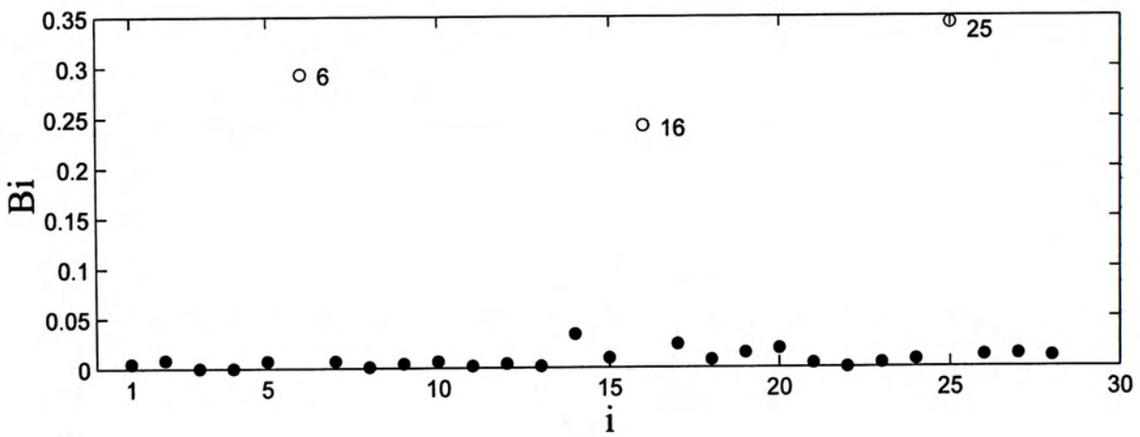

Figure 4.2: Index plot for the brain and body weight data using Method II and the robust estimate

Example 3: Stack loss data set

The index plot of B_i's is given in Figure 4.3. Observations 2, 1, 3 and 17 are declared as outliers. Three more observations 2, 1, 3 are detected as outliers by using the robust estimate while observation 17 is identified as the only outlier by using the sample covariance matrix as shown in Figures 4.3 and 3.3(b) respectively. The three observations 2, 1, 3 have also been detected as outliers by Hadi (1992). The value of B_i for observation 17 by using the revised procedure is less than that by using the proposed procedure in Figure 3.3(b) but it is among the first five largest values (observations 2, 1, 3, 17, 21) as shown in Figure 4.3. Although observation 21 which has been declared as an outlier by Hadi (1992) is not detected as an outlier for the two proposed procedures, its value of B_i is large comparing to the values of B_i for other observations. Therefore, in this example, the revised procedure using the sample robust covariance matrix

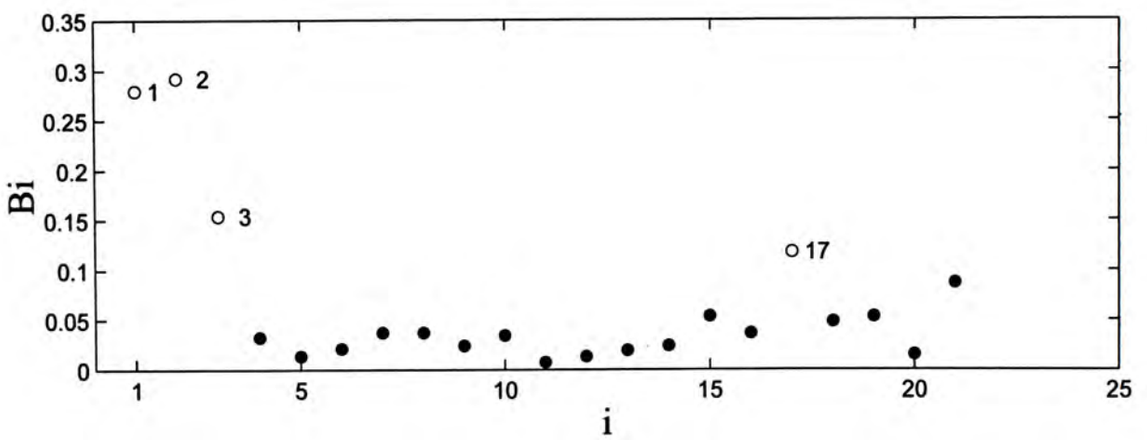

Figure 4.3: Index plot for the stack loss data using Method II and the robust estimate

seems more effective than the procedure proposed in Chapter 3 using the sample covariance matrix.

Example 4: An artificial data set

The index plot in Figure 4.4 shows that observation 9 is the only outlier. The outlier detected is the same as that obtained by using the procedure with sample covariance matrix as shown in Figure 3.4(b). The values of B_i for both the proposed procedure in Chapter 3 and the revised procedure in this chapter are similar as shown in Figures 3.4(b) and 4.4 respectively.

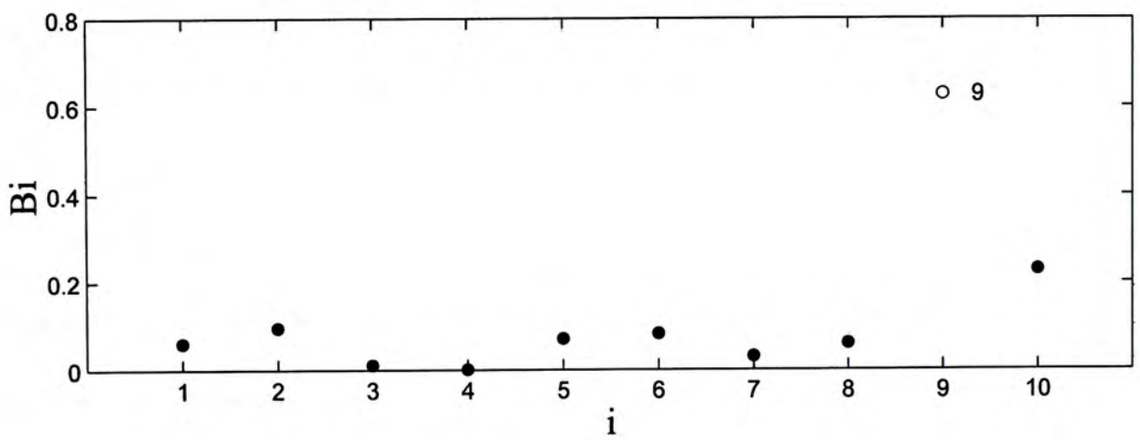

Figure 4.4: Index plot for the artificial data using Method II and the robust estimate

Example 5: An artificial multivariate normal data set

Figure 4.5 gives the index plot of B_i's by using the revised procedure. The last 10 observations, which are constructed to be outliers, are identified as outliers.

Figures 3.5(b) and 4.5 show that the performance of the two procedures using the sample covariance matrix or the sample robust covariance matrix is similar.

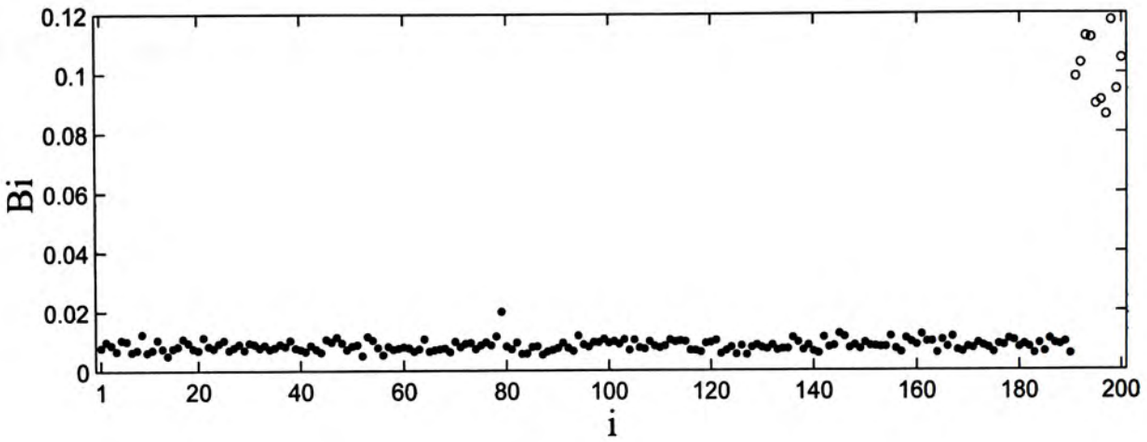

Figure 4.5: Index plot for the artificial multivariate normal data using Method II and the robust estimate

The performance of the procedures using the robust estimate or the sample covariance matrix is similar as shown by the above examples. Therefore, simulation study is carried out to see whether the revised procedure is more effective than the procedure proposed in Chapter 3.

4.3 Simulation Study

The simulation study in this chapter resembles that in the last chapter. All the factor values n, p, d, ε are the same as those values used in the previous simulation study so that the results from the two simulation studies can be compared.

4.3.1 Procedure

The procedure of this simulation study is similar to the one in Section 3.3.2. The proposed procedure in step 6 is replaced by the revised procedure proposed in this chapter and the other steps are the same as before. Here are the steps of the procedure:

1. Choose

$$n = 100, 200, 400, 800 \qquad \text{and}$$

$$p = 5, 10, 20.$$

2. Choose

$$\varepsilon = 0.2, 0.3, 0.4.$$

3. Report

   (a) total number of outliers generated $= n\varepsilon$, and

   (b) total number of good points generated $= n(1-\varepsilon)$.

4. (a) Use $d = 4$ to get far outliers.

   (b) Use $d = 2$ to get close outliers.

5. Draw $n(1-\varepsilon)$ good data points from $N(\mu_0, \Sigma_0) = N(0, I)$ and $n\varepsilon$ bad data points from $N(\mu_0 + \mu, \lambda\Sigma_0) = N(dQ_p^*, I)$ where $Q_p^* = \sqrt{\chi^2_{p,0.999}/p}$.

6. Apply the Method II of the revised procedure to the above dataset.

7. Report

(a) number of data points detected as outliers,

(b) number of good points detected as outliers, and

(c) number of outliers detected without misclassification

= number of data points detected as outliers – number of good points detected as outliers.

8. Calculate the misclassification rate MR_i and the success rate without misclassification SR_i for each trial $i, i = 1, \ldots, 50$, where

$$MR_i = \frac{\text{number of good points detected as outliers in trial } i}{\text{total number of good points generated}}$$

and

$$SR_i = \frac{\text{number of outliers detected without misclassification in trial } i}{\text{total number of outliers generated}}.$$

9. Repeat steps 5–8 for 50 trials for fixed n, p, ε and d.

10. Calculate the average misclassification rate AMR and the average success rate without misclassification ASR, where

$$AMR = \frac{\sum_{i=1}^{50} MR_i}{50}$$

and

$$ASR = \frac{\sum_{i=1}^{50} SR_i}{50}.$$

4.3.2 Results

Similar to the preceding simulation study, the results are also presented in four tables, column (b) of Table 1 to column (b) of Table 4, in the Appendix. The criterion for highlighting the ASR's is the same as the criterion used in the previous simulation study in Chapter 3.

Effect of Dimension

Column (b) of Table 1 reports the AMR's and the ASR's for the revised procedure when the effect of dimension p is considered. Similar to the results given in column (a) of Table 1, most of the highlighted ASR's decrease when p increases except for the underlined ones. The underlined ASR's under column (b) are the same as those underlined values under column (a). We believe that the random effects in generating the data sets cause the underlined ASR's not showing a decreasing trend with increasing dimension. Column (b) of Table 1 also shows that the highlighted ASR's by using the robust estimate are larger than those by using the sample covariance matrix. Moreover, the AMR's for all the highlighted cases are small and are less than the values in column (a) by using the procedure proposed in Chapter 3. The results of the effect of dimension on the revised procedure are better than the results found by using the procedure with sample covariance matrix.

Effect of Sample Size

The effect of sample size n by using the revised procedure is given in column (b) of Table 2. The results resemble those in column (a) of Table 2. The highlighted ASR's in column (b) increase and the corresponding AMR's decrease when sample size increases, except for the underlined ones. Similar to the results in column (a), the underlined ASR's in column (b) are all large (> 0.99) and the corresponding AMR's are zero correcting to the 6 decimal places. Therefore, the underlined ASR's are considered to be satisfactory. Hence, all the highlighted ASR's match with our expectation that ASR increases when sample size n increases. Also, the highlighted ASR's in column (b) are greater than those in column (a) and the corresponding AMR's in column (b) are smaller than those in column (a), indicating that the revised procedure gives better performance.

Effect of Amount of Shift

Column (b) of Table 3 gives the results of the revised procedure when the effect of amount of shift μ or the constant d related to the amount of shift is considered. As expected, all the highlighted ASR's for the far outliers with $d = 4$ are greater than the corresponding ASR's for the close outliers with $d = 2$ when n, p and ε are fixed. The AMR's for the highlighted cases are very small and the largest one is 0.000063. The highlighted results are exactly what we expected and are better than or equal to the corresponding results by using the sample covariance matrix under column (a) of Table 3.

Effect of Fraction of Contamination

The results of the effect of fraction of contamination ε on the revised procedure are reported under column (b) of Table 4. The highlighted results match with our expectation that when the fraction of contamination increases, the ASR decreases. The ASR's found by using the robust estimate in column (b) are larger than those found by using the sample covariance matrix in column (a) while the AMR's in column (b) have smaller values than those in column (a). The revised procedure gives better results when the effect of fraction of contamination is considered.

From the above outcomes, similar results are found for the effect of dimension p, sample size n, fraction of contamination ε and amount of shift μ or the constant d related to μ. The average success rates without misclassification by using the revised procedure are greater than those by adopting the procedure mentioned in Chapter 3. The average misclassification rates by adopting the revised procedure are less than those by using the procedure given in Chapter 3. Therefore, the revised procedure using the robust estimate gives better performance than the procedure proposed in Chapter 3.

Chapter 5

The New Procedure with Random Initial Subset

Atkinson (1994) points out that if the main task is to detect multiple outliers, exact calculation of robust estimates is not needed. The outliers will be found if the basic subset of the algorithm is outlier free. In order to examine whether an outlier free initial subset is crucial to our proposed procedure in Chapter 3, the random initial subset and the volume of the ellipsoid of Atkinson's (1994) algorithm are adopted to modify our procedure in this chapter. The ideas of the random initial subset and the volume of the ellipsoid are given in the first section. The modification of the proposed procedure is proposed in the second section. In the third section, examples are used to illustrate the modified procedure.

5.1 The Elements

Atkinson's (1994) algorithm is similar to the one of Hadi's (1992). Atkinson (1994) uses the random initial subset to replace the initial basic subset of Hadi's (1992). The random initial subset is a subset of p points chosen randomly from a data set with sample size n where p is the dimension of the data. Then, the subset is used to calculate the squared Mahalanobis distance which is the outlier measure used by Atkinson (1994). Randomly chosen initial subsets are repeatedly searched. The best result with the most appropriate random initial subset is determined by the smallest volume of the ellipsoid among the searches. For each search, the volume of the ellipsoid is calculated for each iteration of Step 2 of the procedure. The minimum volume of ellipsoid among the iterations for each search is found. The smallest volume of ellipsoid is the smallest value of the minimum volumes of ellipsoid among the searches.

Procedure

The details of Atkinson's (1994) algorithm with random initial subset are as follows:

Step 1

Choose a subset of p points randomly from a data set with sample size n, where p is the dimension of the data.

Step 2

1. Use the subset of r points from the previous step to calculate the squared Mahalanobis distance $d_i^2(r)$ for observation i, where

$$d_i^2(r) = (x_i - C_r)^T S_r^{-1}(x_i - C_r) \qquad i = 1, \ldots, n$$

where C_r is the $p \times 1$ vector of mean of the subset, S_r is the sample covariance matrix of the subset and x_i is the vector

$$\begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}, \qquad i = 1, \ldots, n$$

where x_{ij} is the ith observation with jth dimension, $j = 1, \ldots, p$.

2. Order $d_i^2(r)$ in ascending order and the observations according to $d_i^2(r)$.

3. Calculate the volume of the ellipsoid $v(r)$ where

$$v(r) = \{|S_r| d_{[med]}^2(r)\}^{\frac{1}{2}}$$

where $d_{[med]}^2(r)$ is the median of the squared Mahalanobis distances $d_i^2(r), i = 1, \ldots, n$.

Step 3

1. The cutoff value to show an observation is an outlier or not is the maximum expected value from a sample of n chi-squared random variables on p degrees

of freedom, approximated by

$$\mathrm{E}(\max \chi_p^2) = \chi^2_{p,(n-0.5)/n}.$$

If the Mahalanobis distance for observation i, $i = 1, \ldots, n$, is greater than the cutoff value, the observation is declared as a suspected outlier.

2. If $r < n$, choose the first $r + 1$ observations from the sorted observations in Step 2 to form a subset and go to Step 2; otherwise, go to Step 4.

Step 4

Choose the minimum value of $v(r)$, $r = p, \ldots, n$, and denote it by $\tilde{v}_a$ where $a = 1, \ldots, d$ and a is the index for searches and d is the total number of searches, for example, Atkinson (1994) uses $d = 100$.

Step 5

1. Repeat Steps 1 to 4 d times and get d searches.

2. The search with the smallest $\tilde{v}_a$ gives the best solution.

3. Plot the stalactite plot of the best solution.

5.2 Procedure

The proposed procedure is modified by using the random initial subset and the volume of the ellipsoid. The initial basic subset of the proposed procedure

is replaced by the random initial subset. The volume of the ellipsoid is used to find the best solution and is calculated in Step 1 of the modified procedure. The steps of the modified procedure using Method II are given below:

Step 0

1. Generate p observation numbers randomly from a discrete uniform distribution $U(1, n)$ without replacement, where p is the dimension of the data and n is the sample size.

2. Regard those p observations as the observations in the basic subset and the other $n - p$ observations as the observations in the non-basic subset.

3. Go to Step 3 and check the stopping criterion before go to Steps 1 and 2.

Step 1

1. Compute the sample mean C_r, the sample covariance matrix S_r of the basic subset and B_i according to the following two cases:

   Case 1: Basic Subset of Full Rank

   If the sample covariance matrix S_r of the basic subset is of full rank, compute

$$B_i(full) = \frac{(x_i - C_r)^T S_r^{-1}(x_i - C_r)}{\sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n}\left[(x_k - C_r)^T S_r^{-1}(x_l - C_r)\right]^2}}, \qquad i = 1, \ldots, n. \quad (5.1)$$

Case 2: Basic Subset Not of Full Rank

(a) If S_r is not of full rank, compute the eigenvalues of S_r, $\lambda_{r1} \geq \ldots \geq \lambda_{rp} = 0$, and the corresponding normalized eigenvectors matrix V_r.

(b) Compute

$$B_i(not\ full) = \frac{(x_i - C_r)^T V_r W_r V_r^T (x_i - C_r)}{\sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n}\left[(x_k - C_r)^T V_r W_r V_r^T (x_l - C_r)\right]^2}}, \quad i = 1, \ldots, n \tag{5.2}$$

where W_r is a diagonal matrix with jth diagonal element,

$$w_{rj} = \frac{1}{\max\{\lambda_{rj}, \lambda_{rs}\}}, \qquad j = 1, \ldots, p$$

and λ_{rs} is the smallest non-zero eigenvalue of S_r.

2. Calculate the volume of the ellipsoid $v(r)$ according to the following two cases:

Case 1: Basic Subset of Full Rank

If S_r is of full rank, calculate

$$v(r) = \{|S_r| B_{med}(full)\}^{\frac{1}{2}} \tag{5.3}$$

where $|S_r|$ is the absolute value of the determinant of S_r,

$$med = \left[\frac{n}{2}\right]$$

is the integer part of $n/2$, $B_{med}(full)$ is the medth $B_i(full)$ (see equation (5.1)) and r is the number of observations in the basic subset.

Case 2: Basic Subset Not of Full Rank

If S_r is not of full rank, calculate

$$v(r) = \{|(V_r W_r V_r^T)^{-1}|B_{med}(not\ full)\}^{\frac{1}{2}} \tag{5.4}$$

where $B_{med}(not\ full)$ is the medth $B_i(not\ full)$ (see equation (5.2)).

Step 2: Increase Size of Basic Subset

1. Rearrange the observations in ascending order according to equation (5.1) or (5.2) depending on whether S_r is of full rank or not.

2. Let r be the number of observations in the current basic subset.

   Divide the observations into two subsets:

   - a basic subset containing the first $r+1$ observations, and
   - a non-basic subset containing the remaining $n-r-1$ observations.

3. Go to Step 3.

Step 3: Stopping Criterion

Method II is used as it gives reasonable results in the previous chapters.

Method II

Case 1: All Observations of Full Rank

If the sample covariance matrix S_r of the basic subset is of full rank, compute

$$b = \frac{\sum_{j=1}^{n}(x_j - C_r)^T S_r^{-1}(x_j - C_r)}{n\sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n}\left[(x_k - C_r)^T S_r^{-1}(x_l - C_r)\right]^2}}, \qquad \forall x_j, x_k, x_l \in \text{set of all observations}$$

and

$$B_i(r) = \frac{(x_i - C_r)^T S_r^{-1}(x_i - C_r)}{\sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n}\left[(x_k - C_r)^T S_r^{-1}(x_l - C_r)\right]^2}}, \quad \begin{array}{l} \forall x_i \in \text{non-basic subset}, \\ \forall x_k, x_l \in \text{set of all observations}, \\ i = 1, \ldots, n-r, \end{array}$$

where r is the number of observations in the current basic subset, C_r is the mean of the basic subset with r observations and S_r is the sample covariance matrix of the basic subset with r observations.

Case 2: All Observations Not of Full Rank

1. If S_r is not of full rank, compute the eigenvalues of S_r, $\lambda_{r1} \geq \ldots \geq \lambda_{rp} = 0$, and the corresponding normalized eigenvectors matrix V_r.

2. Compute

$$b = \frac{\sum_{j=1}^{n}(x_j - C_r)^T V_r W_r V_r^T (x_j - C_r)}{n\sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n}\left[(x_k - C_r)^T V_r W_r V_r^T (x_l - C_r)\right]^2}} ,$$

$\forall x_j, x_k, x_l \in$ set of all observations

and

$$B_i(r) = \frac{(x_i - C_r)^T V_r W_r V_r^T (x_i - C_r)}{\sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n}\left[(x_k - C_r)^T V_r W_r V_r^T (x_l - C_r)\right]^2}} ,$$

$\forall x_i \in$ non-basic subset,

$\forall x_k, x_l \in$ set of all observations,

$i = 1, \ldots, n - r,$

where W_r is a diagonal matrix with jth diagonal element,

$$w_{rj} = \frac{1}{\max\{\lambda_{rj}, \lambda_{rs}\}} , \qquad j = 1, \ldots, p$$

and λ_{rs} is the smallest non-zero eigenvalue of S_r.

Stopping Criterion

1. If the smallest $B_i(r)$ is greater than $2b$, i.e.

$$\text{Min } B_i(r) > 2b , \qquad \forall i \in \text{non-basic subset},$$

regard r as the number of observations in the final basic subset and the observations in the final non-basic subset as the outliers.

2. If $r < n$, go to Steps 1 and 2; otherwise, calculate the volume of ellipsoid $v(n)$ when $r = n$ by using equation (5.3) or (5.4) depending on whether S_r is of full rank or not and go to Step 4.

Step 4

Choose the minimum volume of ellipsoid from $v(r)$'s, $r = p, \ldots, n$, and denote it by $\tilde{v}_a$ where $a = 1, \ldots, 100$ is the index for searches.

Step 5

1. Repeat the Steps 0 to 3 100 times and get 100 searches.

2. Choose the search with the smallest $\tilde{v}_a$. The smaller the $\tilde{v}_a$, the better the search.

3. Regard this search as the best solution.

4. Plot the stalactite plot of the best solution.

5.3 Examples

The first three data sets used in the previous chapters are used to assess the performance of the modified procedure. The artificial data set from Poon, Lew & Poon (1999) is not used as it contains only 10 observations which are too few to obtain random initial subsets for 100 searches. Besides, the artificial multivariate normal data set is ignored because it is computationally expensive to use the data

set for 100 searches. The results are shown in Tables 5.1 to 5.3.

Stalactite plot is used to show the pattern of outliers of the best solution with the smallest volume of ellipsoid. Stalactite plot is a graphical representation of the pattern of outliers when the number of observations r in the basic subset increases from $p+1$ to n where p is the number of observations in the initial basic subset and n is the total number of observations in the data set. The x-axis of the stalactite plot shows the observation numbers and the y-axis shows the number of observations r in the basic subset. The star "$*$" indicates a suspected outlier when the number of observations in the basic subset equals r.

Whether an outlier is stable or not is also shown on the stalactite plot. If an observation is a stable outlier, stars will be shown on the stalactite plot at every successive r from $r = p+1$ to some larger values of r for the corresponding observation. If the observation is not a stable outlier, stars may not appear for some r in between the first star and the final star. That is, the observation is not considered as an outlier for some r. Besides, masking effect can be seen from the stalactite plot when r approaches n as the basic subset includes all the good observations and also outliers.

Different observations are declared as outliers at different values of r. Atkinson & Mulira (1993) suggest that the stalactite plot is most informative when the number of observations r in the basic subset equals 80% or 90% of the sample

size n. The values of r at which outliers are detected in the following examples are between 80% and 90% of the sample size. The dotted lines on the stalactite plots indicate the value of r at which the outliers are detected. Here are the examples:

Example 1: Hawkins, Bradu and Kass data set

The volumes of the ellipsoid for Hawkins, Bradu and Kass data are shown in

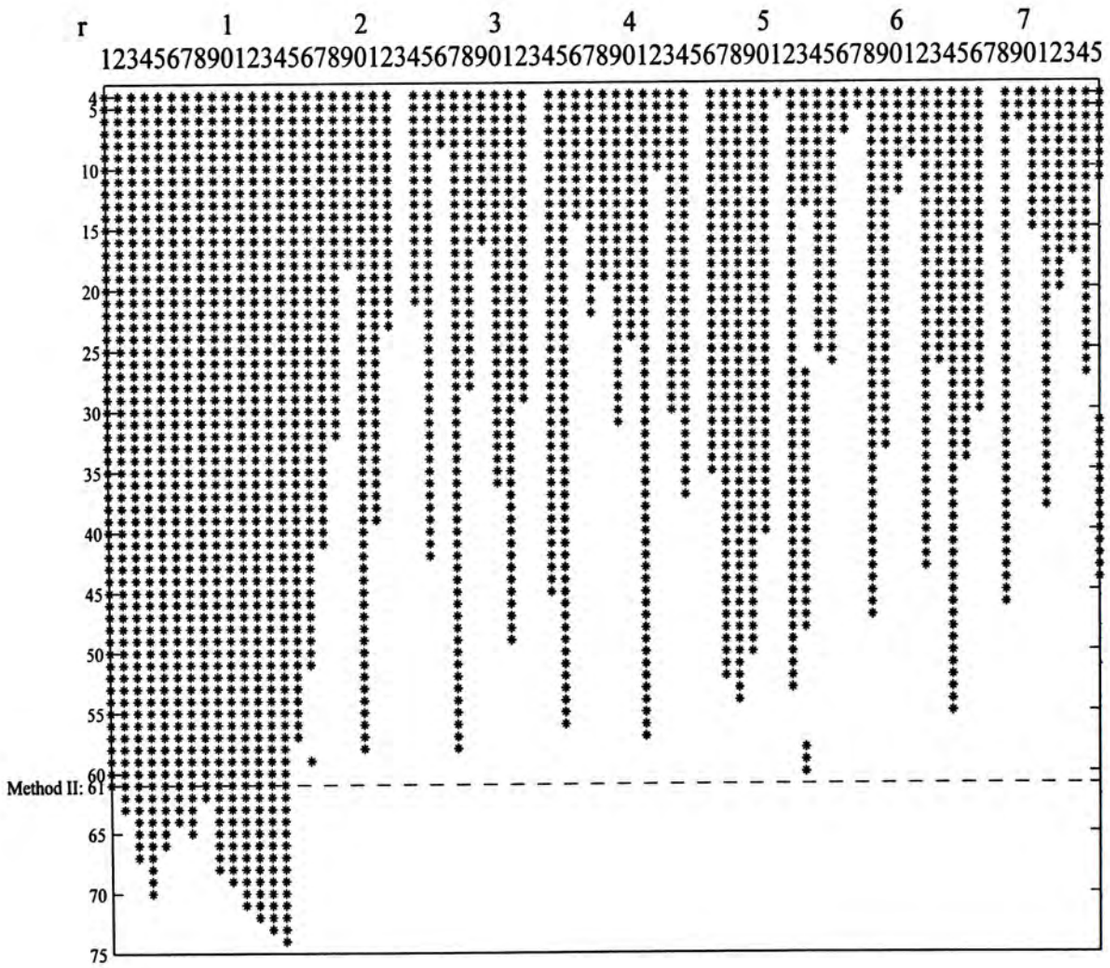

Figure 5.1: Stalactite plot for the Hawkins, Bradu and Kass data using random initial subset with the smallest volume of ellipsoid

Table 5.1. The 77th search with observations 67, 23, 45 as the random initial subset gives the best solution with 0.000067 as the smallest volume of ellipsoid and all the 14 outliers are detected. Other searches do not produce the smallest volume of ellipsoid but all the first 14 observations are also classified as outliers in all the searches. Moreover, the searches with random initial subsets containing one or more of the expected outliers, namely observations 1 to 14, also give the 14 outliers. Figure 5.1 gives the stalactite plot for the 77th search. The 14 stable outliers are declared when the number of observations in the basic subset equals 61 as shown in Figure 5.1. The results are consistent with those in the previous chapters.

Example 2: Brain and body weight data set

Table 5.2 gives the volumes of the ellipsoid for the brain and body weight data. The smallest volume of ellipsoid 0.000393 is found at the 5th search and the 69th search. Observations 23 and 14 are in the random initial subset for the 5th search and observations 24 and 8 are in the random initial subset for the 69th search. If more decimal places are taken up to 10, the 5th and the 69th searches still give the same smallest volume of ellipsoid 0.0003928544. Therefore, both searches give the best solution. Table 5.2 also shows that whether the random initial subsets for all the 100 searches include one or more of the three expected outliers 16, 6, 25 or not does not affect the identification of the three outliers. The stalactite plots for the 5th search and the 69th search are shown in Figures 5.2 and 5.3 respectively. The patterns of outliers are the same for the two searches

regardless the different observations in the random initial subsets. Observations 16, 6, 25 are detected as stable outliers when r equals 25 for both searches. The outliers found are the same as those in the previous chapters.

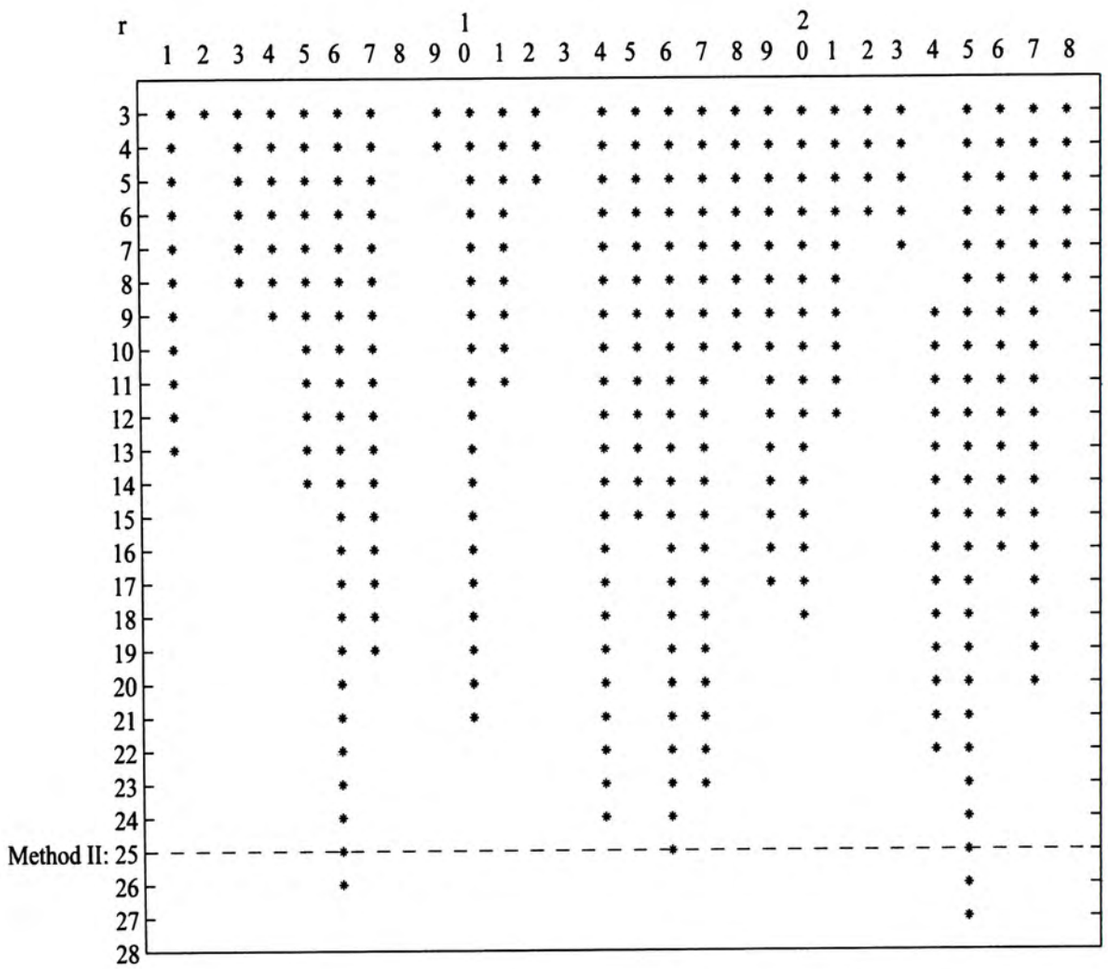

Figure 5.2: Stalactite plot for the brain and body weight data using the random initial subset with the smallest volume of ellipsoid for the 5th search

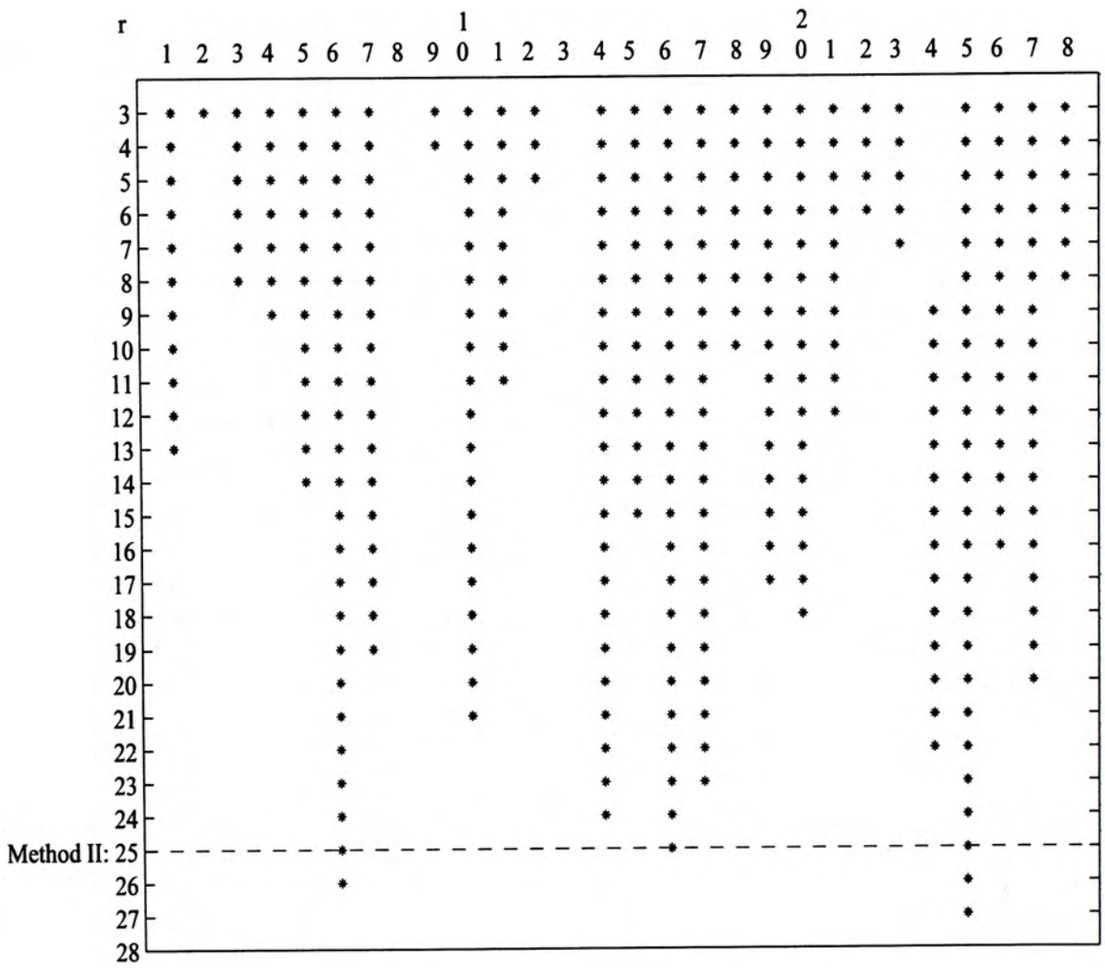

Figure 5.3: Stalactite plot for the brain and body weight data using random initial subset with the smallest volume of ellipsoid for the 69th search

Example 3: Stack loss data set

Table 5.3 shows the volumes of the ellipsoid for the stack loss data. Three types of outliers are found as shown in Table 5.3. The first type of outliers includes observations 3, 1 and 2 with 53 searches. Observations 17 and 21 form the second and the third types of outliers respectively. There are 44 and 3 searches for the second and the third types of outliers. The results in the previous chapters show that the five observations 3, 1, 2, 17 and 21 are far away from the bulk of data.

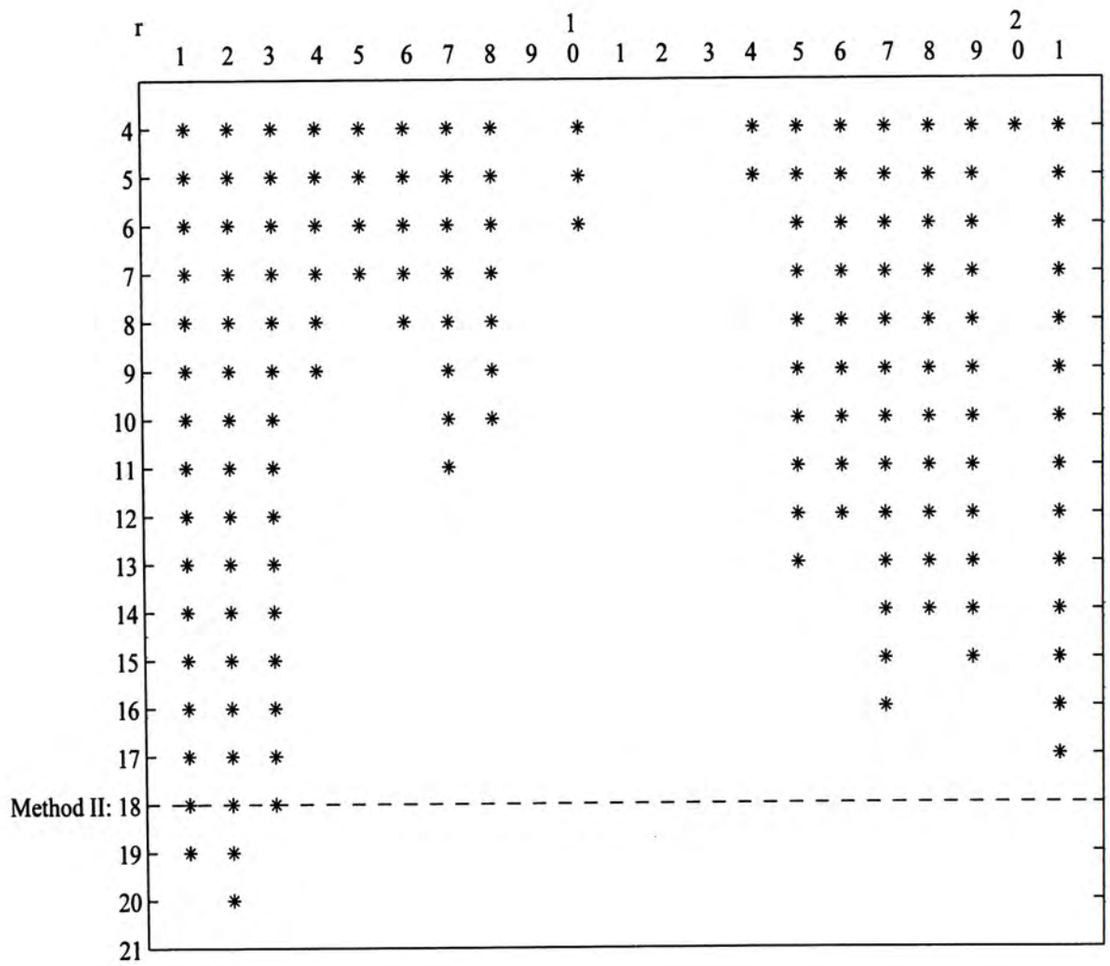

Figure 5.4: Stalactite plot for the stack loss data using random initial subset with smallest volume of ellipsoid

Moreover, the observations are detected as most outlying in all searches regardless the outliers in the random initial subsets. The best solution is found at the 86th search with 0.007726 as the smallest volume of ellipsoid. The random initial subset of the best solution contains observations 13, 12 and 9. Figure 5.4 gives the stalactite plot for the 86th search. Observations 2, 1 and 3 are identified as stable outliers when r equals 18.

From the above examples, whether the random initial subsets contain the expected outliers or not does not affect the performance of the procedure proposed in Chapter 3. Therefore, an outlier free initial subset is not critical to the procedure mentioned in Chapter 3.

Table 5.1: The volume of the ellipsoid for Hawkins, Bradu and Kass data

Search number	Initial observations	$\tilde{v}_a$	Outliers
1	20 58 43	0.000639	1 8 2 6 7 10 3 9 5 4 11 13 12 14
2	30 40 3	0.001291	1 8 2 6 7 10 3 9 5 4 11 13 12 14
3	40 21 37	0.000558	1 8 2 6 7 10 3 9 5 4 11 13 12 14
4	50 3 71	0.008593	1 8 2 6 7 10 3 9 5 4 11 13 12 14
5	60 35 31	0.010253	1 8 2 6 7 10 3 9 5 4 11 13 12 14
6	44 25 12	0.001291	1 8 2 6 7 10 3 9 5 4 11 13 12 14
7	54 6 47	0.003744	1 8 2 6 7 10 3 9 5 4 11 13 12 14
8	63 39 6	0.003742	1 8 2 6 7 10 3 9 5 4 11 13 12 14
9	48 28 62	0.000687	1 8 2 6 7 10 3 9 5 4 11 13 12 14
10	57 10 22	0.008866	1 8 2 6 7 10 3 9 5 4 11 13 12 14
11	67 66 56	0.000286	1 8 2 6 7 10 3 9 5 4 11 13 12 14
12	2 48 16	0.003630	1 8 2 6 7 10 3 9 5 4 11 13 12 14
13	12 30 50	0.001983	1 8 2 6 7 10 3 9 5 4 11 13 12 14
14	22 11 9	0.001291	1 8 2 6 7 10 3 9 5 4 11 13 12 14
15	32 68 44	0.000891	1 8 2 6 7 10 3 9 5 4 11 13 12 14
16	42 50 3	0.002378	1 8 2 6 7 10 3 9 5 4 11 13 12 14
17	52 31 38	0.000132	1 8 2 6 7 10 3 9 5 4 11 13 12 14
18	61 13 72	0.008366	1 8 2 6 7 10 3 9 5 4 11 13 12 14
19	71 70 31	0.017074	1 8 2 6 7 10 3 9 5 4 11 13 12 14
20	6 51 66	0.007862	1 8 2 6 7 10 3 9 5 4 11 13 12 14
21	16 33 25	0.000546	1 8 2 6 7 10 3 9 5 4 11 13 12 14
22	26 15 60	0.017074	1 8 2 6 7 10 3 9 5 4 11 13 12 14
23	36 71 19	0.000108	1 8 2 6 7 10 3 9 5 4 11 13 12 14
24	46 53 21	0.000124	1 8 2 6 7 10 3 9 5 4 11 13 12 14
25	30 18 35	0.000819	1 8 2 6 7 10 3 9 5 4 11 13 12 14
26	40 75 69	0.000225	1 8 2 6 7 10 3 9 5 4 11 13 12 14
27	50 56 29	0.001628	1 8 2 6 7 10 3 9 5 4 11 13 12 14
28	59 38 63	0.000656	1 8 2 6 7 10 3 9 5 4 11 13 12 14
29	69 20 22	0.000641	1 8 2 6 7 10 3 9 5 4 11 13 12 14
30	4 1 57	0.003534	1 8 2 6 7 10 3 9 5 4 11 13 12 14
31	14 58 16	0.017074	1 8 2 6 7 10 3 9 5 4 11 13 12 14
32	24 40 51	0.017074	1 8 2 6 7 10 3 9 5 4 11 13 12 14
33	34 21 10	0.002272	1 8 2 6 7 10 3 9 5 4 11 13 12 14
34	44 3 19	0.009273	1 8 2 6 7 10 3 9 5 4 11 13 12 14
35	28 43 26	0.017074	1 8 2 6 7 10 3 9 5 4 11 13 12 14
36	38 25 60	0.000441	1 8 2 6 7 10 3 9 5 4 11 13 12 14
37	48 7 20	0.009170	1 8 2 6 7 10 3 9 5 4 11 13 12 14
38	57 63 54	0.017074	1 8 2 6 7 10 3 9 5 4 11 13 12 14
39	67 45 13	0.008910	1 8 2 6 7 10 3 9 5 4 11 13 12 14
40	2 27 48	0.012643	1 8 2 6 7 10 3 9 5 4 11 13 12 14
41	12 8 7	0.009259	1 8 2 6 7 10 3 9 5 4 11 13 12 14
42	22 65 42	0.017074	1 8 2 6 7 10 3 9 5 4 11 13 12 14
43	32 47 1	0.003744	1 8 2 6 7 10 3 9 5 4 11 13 12 14
44	42 28 35	0.001365	1 8 2 6 7 10 3 9 5 4 11 13 12 14
45	51 10 70	0.006658	1 8 2 6 7 10 3 9 5 4 11 13 12 14
46	61 67 29	0.017074	1 8 2 6 7 10 3 9 5 4 11 13 12 14
47	71 48 64	0.017074	1 8 2 6 7 10 3 9 5 4 11 13 12 14
48	6 30 23	0.009180	1 8 2 6 7 10 3 9 5 4 11 13 12 14
49	16 12 57	0.008333	1 8 2 6 7 10 3 9 5 4 11 13 12 14
50	26 68 17	0.009259	1 8 2 6 7 10 3 9 5 4 11 13 12 14

Table 5.1 (continue): The volume of the ellipsoid for Hawkins, Bradu and Kass data set

Search number	Initial observations	$\tilde{v}_a$	Outliers
51	36 50 51	0.009397	1 8 2 6 7 10 3 9 5 4 11 13 12 14
52	45 32 11	0.007363	1 8 2 6 7 10 3 9 5 4 11 13 12 14
53	55 13 45	0.005766	1 8 2 6 7 10 3 9 5 4 11 13 12 14
54	65 70 4	0.008091	1 8 2 6 7 10 3 9 5 4 11 13 12 14
55	75 52 39	0.000957	1 8 2 6 7 10 3 9 5 4 11 13 12 14
56	10 33 73	0.013455	1 8 2 6 7 10 3 9 5 4 11 13 12 14
57	20 15 33	0.008910	1 8 2 6 7 10 3 9 5 4 11 13 12 14
58	30 72 67	0.017074	1 8 2 6 7 10 3 9 5 4 11 13 12 14
59	40 53 26	0.002737	1 8 2 6 7 10 3 9 5 4 11 13 12 14
60	49 35 61	0.000166	1 8 2 6 7 10 3 9 5 4 11 13 12 14
61	59 17 20	0.017074	1 8 2 6 7 10 3 9 5 4 11 13 12 14
62	69 73 54	0.017074	1 8 2 6 7 10 3 9 5 4 11 13 12 14
63	4 55 14	0.001291	1 8 2 6 7 10 3 9 5 4 11 13 12 14
64	14 37 48	0.017074	1 8 2 6 7 10 3 9 5 4 11 13 12 14
65	24 18 8	0.003744	1 8 2 6 7 10 3 9 5 4 11 13 12 14
66	34 75 42	0.017074	1 8 2 6 7 10 3 9 5 4 11 13 12 14
67	43 57 1	0.003744	1 8 2 6 7 10 3 9 5 4 11 13 12 14
68	53 38 36	0.000574	1 8 2 6 7 10 3 9 5 4 11 13 12 14
69	63 20 70	0.017074	1 8 2 6 7 10 3 9 5 4 11 13 12 14
70	73 2 30	0.003918	1 8 2 6 7 10 3 9 5 4 11 13 12 14
71	8 58 64	0.009259	1 8 2 6 7 10 3 9 5 4 11 13 12 14
72	18 40 23	0.017074	1 8 2 6 7 10 3 9 5 4 11 13 12 14
73	28 22 58	0.000266	1 8 2 6 7 10 3 9 5 4 11 13 12 14
74	38 3 17	0.000558	1 8 2 6 7 10 3 9 5 4 11 13 12 14
75	47 60 52	0.000565	1 8 2 6 7 10 3 9 5 4 11 13 12 14
76	57 42 11	0.007583	1 8 2 6 7 10 3 9 5 4 11 13 12 14
77	67 23 45	**0.000067**	1 8 2 6 7 10 3 9 5 4 11 13 12 14
78	2 5 52	0.009259	1 8 2 6 7 10 3 9 5 4 11 13 12 14
79	61 45 37	0.000773	1 8 2 6 7 10 3 9 5 4 11 13 12 14
80	45 10 43	0.003514	1 8 2 6 7 10 3 9 5 4 11 13 12 14
81	55 67 2	0.006512	1 8 2 6 7 10 3 9 5 4 11 13 12 14
82	65 48 36	0.017074	1 8 2 6 7 10 3 9 5 4 11 13 12 14
83	75 30 71	0.000262	1 8 2 6 7 10 3 9 5 4 11 13 12 14
84	10 12 30	0.008910	1 8 2 6 7 10 3 9 5 4 11 13 12 14
85	20 68 65	0.017074	1 8 2 6 7 10 3 9 5 4 11 13 12 14
86	30 50 24	0.000490	1 8 2 6 7 10 3 9 5 4 11 13 12 14
87	39 32 58	0.000670	1 8 2 6 7 10 3 9 5 4 11 13 12 14
88	49 13 18	0.007378	1 8 2 6 7 10 3 9 5 4 11 13 12 14
89	59 70 52	0.017074	1 8 2 6 7 10 3 9 5 4 11 13 12 14
90	69 52 12	0.007471	1 8 2 6 7 10 3 9 5 4 11 13 12 14
91	4 33 46	0.009180	1 8 2 6 7 10 3 9 5 4 11 13 12 14
92	14 15 5	0.000883	1 8 2 6 7 10 3 9 5 4 11 13 12 14
93	24 72 40	0.000197	1 8 2 6 7 10 3 9 5 4 11 13 12 14
94	33 53 74	0.000071	1 8 2 6 7 10 3 9 5 4 11 13 12 14
95	43 35 34	0.017074	1 8 2 6 7 10 3 9 5 4 11 13 12 14
96	53 17 68	0.017074	1 8 2 6 7 10 3 9 5 4 11 13 12 14
97	63 73 27	0.017074	1 8 2 6 7 10 3 9 5 4 11 13 12 14
98	73 55 62	0.000216	1 8 2 6 7 10 3 9 5 4 11 13 12 14
99	8 37 21	0.008490	1 8 2 6 7 10 3 9 5 4 11 13 12 14
100	18 68 56	0.000207	1 8 2 6 7 10 3 9 5 4 11 13 12 14

Table 5.2: The volume of the ellipsoid for the brain and body weight data

Search number	Initial observations	$\tilde{v}_a$	Outliers
1	8 22	0.000692	16 6 25
2	11 15	0.001435	16 6 25
3	15 8	0.009790	16 6 25
4	19 1	0.004867	16 6 25
5	23 14	**0.000393**	16 6 25
6	17 10	0.002776	16 6 25
7	20 3	0.000699	16 6 25
8	24 15	0.002227	16 6 25
9	18 11	0.003016	16 6 25
10	22 4	0.000663	16 6 25
11	25 17	0.001845	16 6 25
12	20 12	0.009790	16 6 25
13	23 5	0.001144	16 6 25
14	27 26	0.017308	16 6 25
15	3 20	0.000699	16 6 25
16	6 13	0.002330	16 6 25
17	10 6	0.001327	16 6 25
18	14 27	0.006016	16 6 25
19	17 20	0.000847	16 6 25
20	21 13	0.001532	16 6 25
21	25 6	0.009790	16 6 25
22	28 20	0.004182	16 6 25
23	22 15	0.000692	16 6 25
24	26 8	0.009790	16 6 25
25	2 1	0.002818	16 6 25
26	6 22	0.002330	16 6 25
27	9 15	0.001478	16 6 25
28	13 8	0.000393	16 6 25
29	17 1	0.001163	16 6 25
30	20 23	0.001730	16 6 25
31	24 16	0.005910	16 6 25
32	28 9	0.003585	16 6 25
33	3 2	0.001746	16 6 25
34	7 23	0.003010	16 6 25
35	11 16	0.010134	16 6 25
36	14 10	0.009790	16 6 25
37	18 3	0.010134	16 6 25
38	22 24	0.003029	16 6 25
39	25 17	0.001845	16 6 25
40	1 10	0.002749	16 6 25
41	5 3	0.002749	16 6 25
42	8 25	0.003636	16 6 25
43	12 18	0.009790	16 6 25
44	16 11	0.010134	16 6 25
45	20 4	0.003234	16 6 25
46	23 25	0.003095	16 6 25
47	27 18	0.007142	16 6 25
48	3 11	0.009790	16 6 25
49	6 5	0.004697	16 6 25
50	10 26	0.009790	16 6 25

Table 5.2 (continue): The volume of the ellipsoid for the brain and body weight data

Search number	Initial observations	$\tilde{v}_a$	Outliers
51	14 19	0.009922	16 6 25
52	17 12	0.003903	16 6 25
53	21 5	0.001500	16 6 25
54	25 26	0.003016	16 6 25
55	28 20	0.004182	16 6 25
56	4 13	0.000663	16 6 25
57	8 6	0.002330	16 6 25
58	11 27	0.009790	16 6 25
59	15 20	0.001767	16 6 25
60	19 13	0.009790	16 6 25
61	22 7	0.009790	16 6 25
62	26 28	0.002749	16 6 25
63	2 21	0.002818	16 6 25
64	5 14	0.009790	16 6 25
65	9 7	0.003585	16 6 25
66	13 28	0.004867	16 6 25
67	17 21	0.009790	16 6 25
68	20 15	0.001767	16 6 25
69	24 8	**0.000393**	16 6 25
70	28 1	0.002040	16 6 25
71	3 22	0.000527	16 6 25
72	7 15	0.009986	16 6 25
73	11 8	0.009790	16 6 25
74	14 2	0.004867	16 6 25
75	18 23	0.009790	16 6 25
76	22 16	0.004385	16 6 25
77	25 9	0.002077	16 6 25
78	1 2	0.002818	16 6 25
79	5 23	0.001144	16 6 25
80	8 16	0.002330	16 6 25
81	12 10	0.009790	16 6 25
82	16 3	0.002330	16 6 25
83	19 24	0.001978	16 6 25
84	23 17	0.000953	16 6 25
85	27 10	0.009790	16 6 25
86	3 22	0.000527	16 6 25
87	25 18	0.010134	16 6 25
88	28 12	0.000393	16 6 25
89	4 5	0.009790	16 6 25
90	8 26	0.009790	16 6 25
91	11 19	0.002049	16 6 25
92	15 12	0.009986	16 6 25
93	19 5	0.000398	16 6 25
94	22 26	0.001459	16 6 25
95	26 20	0.007746	16 6 25
96	2 13	0.000393	16 6 25
97	5 6	0.004697	16 6 25
98	9 27	0.009790	16 6 25
99	13 20	0.009790	16 6 25
100	16 13	0.003771	16 6 25

Table 5.3: The volume of the ellipsoid for the stack loss data

Search number	Initial observations	$\tilde{v}_a$	Outliers
1	6 17 12	0.858573	3 1 2
2	9 11 1	0.158474	17
3	7 18 6	0.709080	3 1 2
4	10 12 15	0.246912	17
5	15 2 13	1.060940	17
6	14 8 18	0.231396	17
7	16 3 6	0.033467	3 1 2
8	19 13 16	0.246912	17
9	15 9 11	0.173507	17
10	18 4 21	0.466194	17
11	20 14 9	0.231396	17
12	16 10 4	1.087694	3 1 2
13	19 5 14	0.396215	17
14	7 12 15	0.216625	3 1 2
15	10 6 3	0.056979	3 1 2
16	13 1 6	0.509585	17
17	6 19 12	0.736790	3 1 2
18	9 13 1	0.935762	3 1 2
19	12 8 10	0.231396	17
20	15 3 20	0.404477	17
21	20 14 18	0.077189	3 1 2
22	2 9 7	1.319813	3 1 2
23	5 4 16	0.376984	3 1 2
24	8 19 5	0.412835	3 1 2
25	10 14 15	0.176645	3 1 2
26	13 9 3	0.287691	3 1 2
27	16 4 13	1.231668	3 1 2
28	19 20 2	0.643775	21
29	21 15 11	0.735967	3 1 2
30	3 10 21	0.692362	3 1 2
31	6 5 9	0.174709	17
32	9 20 19	0.466194	17
33	11 15 8	0.461835	3 1 2
34	14 10 17	0.304741	17
35	17 5 6	0.790319	3 1 2
36	1 16 4	0.068144	17
37	4 11 14	0.104606	3 1 2
38	7 5 3	0.709080	3 1 2
39	10 21 12	0.408637	3 1 2
40	13 16 1	1.590834	3 1 2
41	15 11 10	0.173896	3 1 2
42	18 6 20	2.956912	17
43	3 17 18	0.452818	3 1 2
44	5 12 7	0.280956	21
45	11 1 5	0.120050	17
46	19 7 13	0.251618	17
47	17 13 18	2.114432	21
48	2 3 16	0.428695	17
49	5 19 4	0.466194	17
50	10 8 3	0.935762	3 1 2

Table 5.3 (continue): The volume of the ellipsoid for the stack loss data

Search number	Initial observations	$\tilde{v}_a$	Outliers
51	16 19 1	0.571529	17
52	19 14 11	0.256084	3 1 2
53	21 9 20	0.466194	17
54	6 20 18	2.956912	17
55	9 14 7	0.139081	3 1 2
56	11 9 17	0.231396	17
57	14 4 5	0.226048	3 1 2
58	17 20 15	0.294165	3 1 2
59	20 15 4	0.231396	17
60	4 5 2	0.064207	17
61	7 20 12	0.262863	3 1 2
62	10 15 21	0.176645	3 1 2
63	12 10 6	0.081241	3 1 2
64	11 16 14	0.364437	3 1 2
65	14 11 3	0.315338	17
66	16 6 12	0.110394	17
67	15 12 17	0.246912	17
68	20 2 15	0.087518	17
69	2 18 4	0.709080	3 1 2
70	5 13 20	0.060375	3 1 2
71	3 19 18	0.465832	3 1 2
72	6 13 21	0.373853	3 1 2
73	1 4 16	0.068144	17
74	18 15 17	0.044319	3 1 2
75	3 5 15	0.028246	17
76	5 20 4	0.466194	17
77	8 15 14	1.431335	3 1 2
78	11 10 2	0.275222	3 1 2
79	14 5 12	0.082363	3 1 2
80	16 21 10	0.304741	17
81	12 11 16	0.334976	3 1 2
82	15 6 5	0.461835	3 1 2
83	17 1 15	0.417208	17
84	5 7 1	0.360625	17
85	10 17 21	1.431335	3 1 2
86	13 12 9	**0.007726**	3 1 2
87	16 7 19	0.172019	3 1 2
88	19 2 8	0.709080	3 1 2
89	21 18 17	0.533371	3 1 2
90	3 13 6	0.287691	3 1 2
91	6 7 15	1.142475	17
92	11 18 14	0.256084	3 1 2
93	14 13 2	0.425099	17
94	17 8 12	0.231396	17
95	20 3 1	0.190604	17
96	1 19 10	0.592644	17
97	4 14 20	0.231396	17
98	7 8 9	0.251618	17
99	10 3 18	0.138794	3 1 2
100	15 14 16	2.923257	3 1 2

Chapter 6

Discussion

In this chapter, several aspects from the previous chapters will be discussed. Alternative forms of the outlier measures used in Steps 1 and 2 of the proposed procedure in Chapter 3 and of the revised procedure in Chapter 4 are discussed first. Then, another way of finding the robust version of the sample covariance matrix S_M in Step 0 of the revised procedure in Chapter 4 is given. Next, the problem of Method I and the factor in affecting the identification of outliers by using Method I are described.

Alternative Forms of the Outlier Measures

The purpose of the outlier measures in the Steps 1 and 2 of the procedures mentioned in Chapters 3 (equations (3.1) and (3.2)) and 4 (equations (4.4) and (4.5)) is to rearrange the observations. Whether the outlier measures are standardized by the square root part of the outlier measures or not does not affect the rearrangement of the observations. Therefore, the alternative forms of the

equations become only the numerators of the equations. However, the standardization cannot be neglected in the stopping criterion in Step 3; otherwise, the procedures will lose the nice properties inherited from Poon, Lew & Poon (2000).

Another Way of Finding S_M

Instead of computing S_M in Step 0 of the revised procedure in Chapter 4 by using equation (4.2), we can use Hadi's (1992) approach in Chapter 2 to calculate the robust estimate by adopting a weight function. Only the first two parts of Step 0 of the revised procedure are changed and they are given in the following. The notations used are the same as those in Chapter 4.

1. (a) Compute the co-ordinatewise median vector C_M and S_M which are defined in equations (4.1) and (4.2) respectively.

   (b) Rearrange the observations in ascending order according to the robust distance

$$B_i(C_M, S_M) = \frac{(x_i - C_M)^T S_M^{-1}(x_i - C_M)}{\sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n}\left[(x_k - C_M)^T S_M^{-1}(x_l - C_M)\right]^2}}, \qquad i = 1, \ldots, n.$$

   (c) Define the weight function

$$v_i = \begin{cases} 1, & \text{if } i \leq \text{ integer part of } (n+p+1)/2, \\ 0, & \text{otherwise,} \end{cases} \qquad i = 1, \ldots, n.$$

(d) Compute

$$C_R = \frac{\sum_{i=1}^{n} v_i x_i}{\sum_{i=1}^{n} v_i} \quad \text{and} \quad S_R = \frac{\sum_{i=1}^{n} v_i (x_i - C_R)(x_i - C_R)^T}{\sum_{i=1}^{n} v_i - 1}.$$

2. Rearrange the n observations in ascending order according to

$$B_i(C_R, S_R) = \frac{(x_i - C_R)^T S_R^{-1} (x_i - C_R)}{\sqrt{\sum_{k=1}^{n} \sum_{l=1}^{n} \left[(x_k - C_R)^T S_R^{-1} (x_l - C_R) \right]^2}}, \qquad i = 1, \ldots, n.$$

Since S_M is simpler than S_R, S_M is adopted in Chapter 4. C_R and S_R may be used instead of C_M and S_M in the future research.

Factor Affecting Method I

The factor affecting the detection of outliers by using Method I is b of the benchmark $2b$ in the stopping criterion. The value of b affects the number of iterations of Steps 1 to 2 of the procedure before reaching the stopping criterion and hence the number of observations in the non-basic subset. The value of b is found to be usually very small and the stopping criterion is met very easily. The index plots (Figures 3.1(a) to 3.5(a)) in Chapter 3 show that the number of outliers found by using Method I is greater than half of the observations in the whole data set.

The first three data sets used in the previous chapters and stalactite plots are used to show the above situation. The patterns of outliers shown by the stalac-

tite plots for Method I and Method II should be the same because the stopping criterion will not affect the construction of the stalactite plots. Therefore, only one stalactite plot is shown for both methods. Horizontal solid lines and dotted lines are used to indicate when the procedure stops for Method I and Method II respectively. Two cases, using the proposed procedure with sample covariance matrix and the revised procedure with robust estimate, are also considered. The stalactite plot by using the proposed procedure in Chapter 3 is given first, followed by the stalactite plot by using the revised procedure, in each example. The outliers found in the following examples are the same as those detected in Chapters 3 and 4. The stalactite plots of the examples are shown below:

Example 1: Hawkins, Bradu and Kass data set

The respective stalactite plots for the Hawkins, Bradu and Kass data using the sample covariance matrix and the robust estimate are given in Figures 6.1 and 6.2. Both figures show that observations 1 to 14 are declared as stable outliers when $r = 61$ for Method II. The figures also show that nearly all the observations are detected as outliers when the procedure stops early at $r = 4$ for Method I.

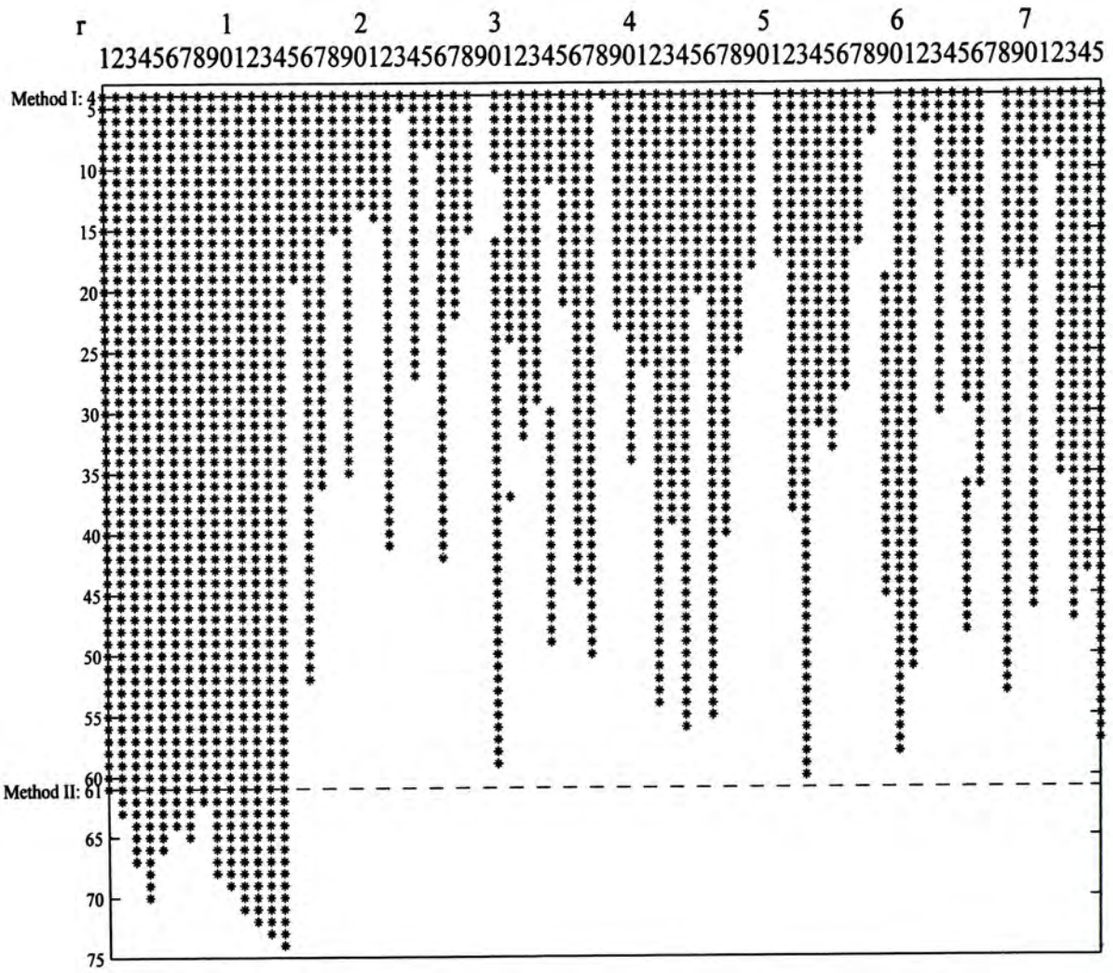

Figure 6.1: Stalactite plot for the Hawkins, Bradu and Kass data using the sample covariance matrix

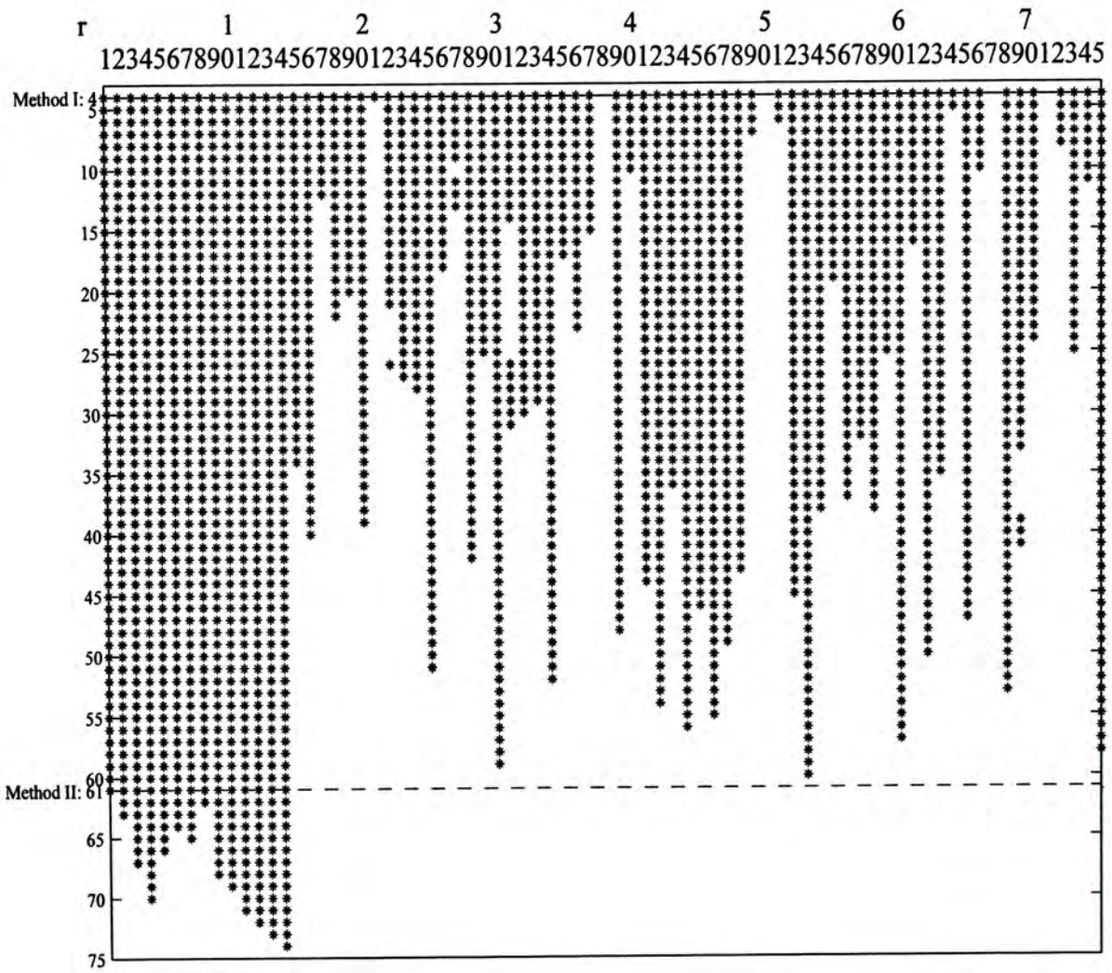

Figure 6.2: Stalactite plot for the Hawkins, Bradu and Kass data using the robust estimate

Example 2: Brain and body weight data set

Figures 6.3 and 6.4 give the stalactite plots for the brain and body weight data using the sample covariance matrix and the sample robust covariance matrix respectively. Three observations 6, 16, 25 are detected as stable outliers when $r = 25$ for Method II in both figures. For Method I, both figures show that the procedure ends at small values of r and more than half of the observations are classified as outliers.

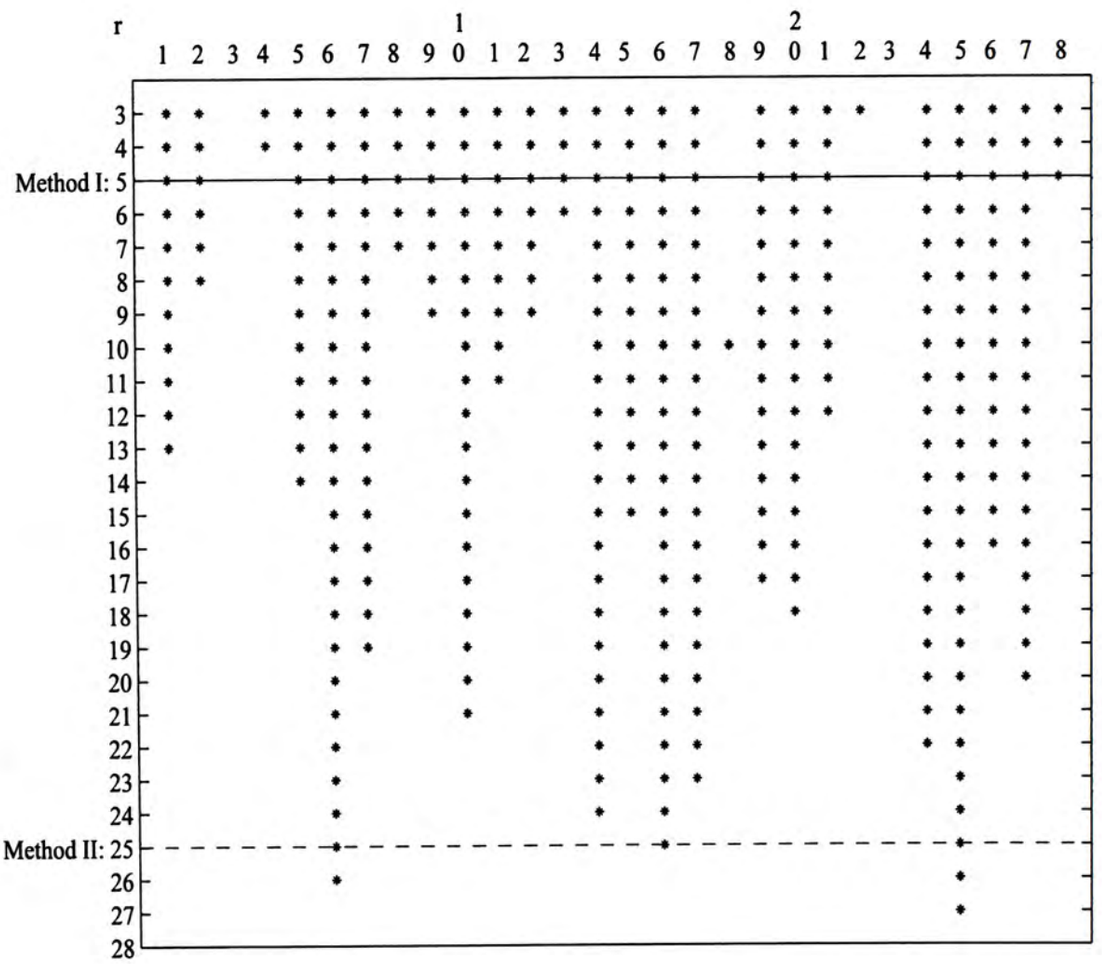

Figure 6.3: Stalactite plot for the brain and body weight data the sample covariance matrix

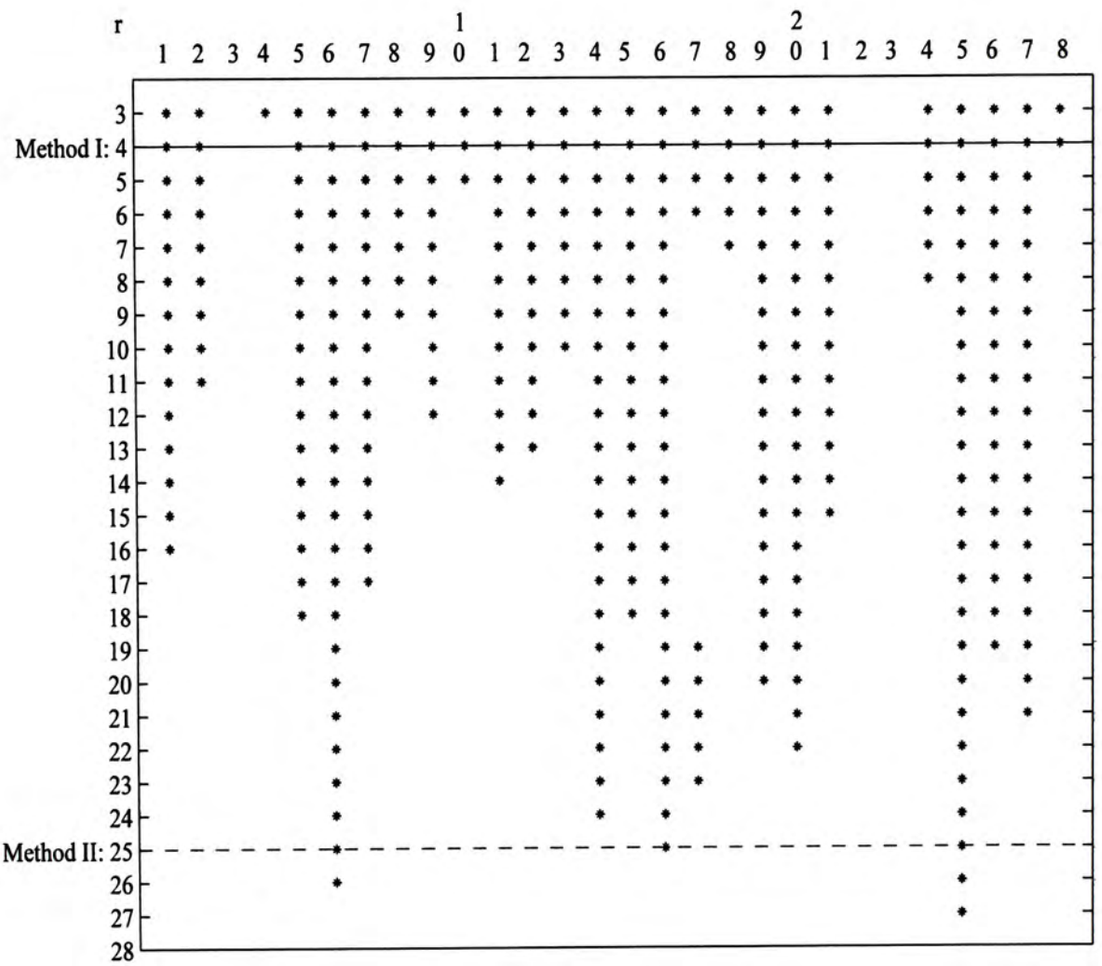

Figure 6.4: Stalactite plot for the brain and body weight data using the robust estimate

Example 3: Stack loss data set

The stalactite plots for the stack loss data using the sample covariance matrix and the robust estimate are shown in Figures 6.5 and 6.6 respectively. Observation 17 is identified as the unstable outlier when $r = 20$ for Method II in Figure 6.5 and more than half of the observations are detected as outliers when $r = 6$ for Method I. Observations 1, 2, 3 and 17 are declared to be stable outliers when r reaches 17 as shown in Figure 6.6. Figure 6.6 also shows that the procedure terminates too early at $r = 4$ and too many outliers are detected.

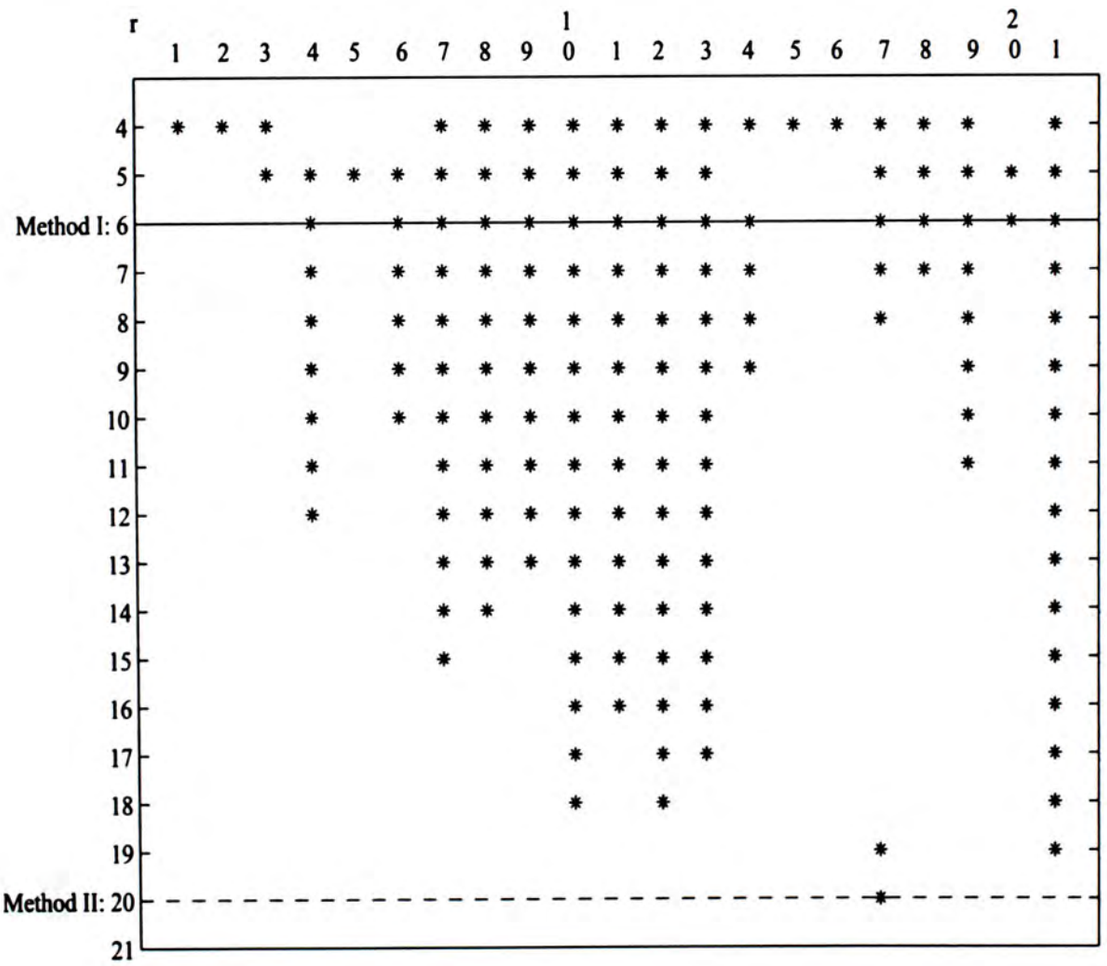

Figure 6.5: Stalactite plot for the stack loss data set using the sample covariance matrix

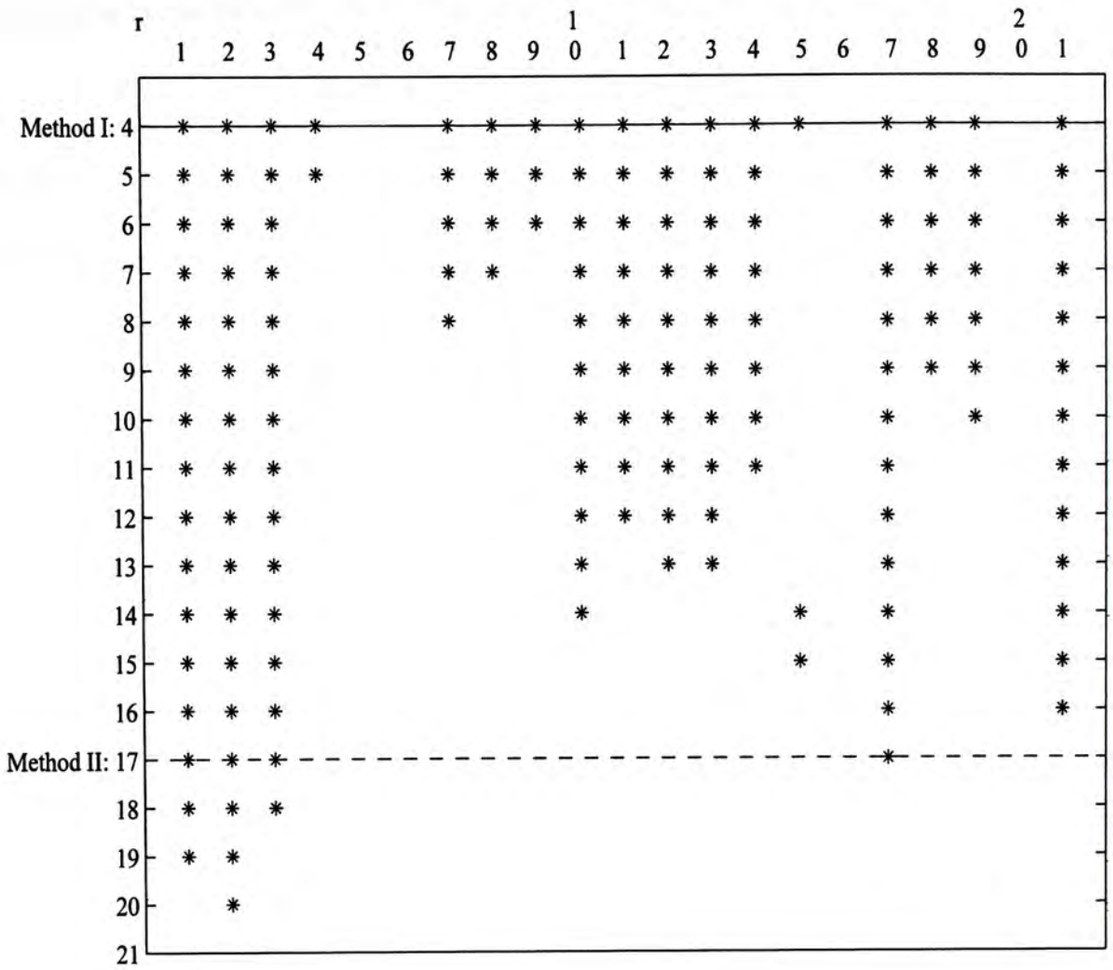

Figure 6.6: Stalactite plot for the stack loss data set using the robust estimate

From the above examples, reasonable number of outliers are identified for Method II and too many outliers are detected at small values of r for Method I. This suggests that the procedure using method I stops too early. This is due to the small value of b of the benchmark. The stopping criterion is met early and the number of observations in the final basic subset is small and hence the number of outliers identified is greater than half of the total number of observations. This leads to the unreasonable number of outliers detected. In order to increase the number of observations in the final basic subset, b should be increased. The idea

of using a correction factor in Chapter 2 may be adopted to increase and adjust the value of b so as to make Method I as effective as Method II in detecting outliers. The form of the correction factor which may depends on various values of the number of observations r in the basic subset, sample size n and dimension p may be found by a simulation study.

Chapter 7

Conclusion

In this thesis, the new stepping procedure based on Hadi's (1992) stepping algorithm and Poon, Lew & Poon's (2000) outlier measure and benchmark is proposed to identify multiple multivariate outliers. Two kinds of metrics and two kinds of methods called Method I and Method II are considered. The two kinds of metrics are the sample covariance matrix and the sample robust covariance matrix. Method I uses the observations in the basic subset and Method II uses all the observations in both the basic subset and the non-basic subset. The new procedure by using both methods with the sample covariance matrix is proposed in Chapter 3. Some reported data sets and artificial data sets are used to assess the performance of the proposed procedure. The index plots of the data sets in Chapter 3 show that the proposed procedure is effective in identifying outliers when Method II is used. The procedure gives unsatisfactory results that more than half of the observations are identified as outliers when Method I is adopted. Therefore, Method I is neglected in the simulation study and in the subsequent

two chapters. Simulation study is performed to examine the performance of the proposed procedure. The average success rate without misclassification decreases in the simulation study when the fraction of contamination increases. On the other hand, the sample covariance matrix is highly affected by outlying observations, so a refinement of the proposed procedure is made in Chapter 4. The revised procedure is similar to the procedure proposed in Chapter 3 but a sample robust covariance matrix is used instead of the sample covariance matrix. The same data sets are applied to the revised procedure. The index plots in Chapter 4 indicate that the revised procedure by using Method II is as effective as the proposed procedure in Chapter 3 in detecting multivariate outliers. In order to compare the performance of the proposed procedure and the revised procedure, a simulation study is performed on the revised procedure. The revised procedure gives larger average success rate without misclassification and smaller average misclassification rate than the proposed procedure. That is, the performance of the new procedure is better when the sample robust covariance matrix and Method II are adopted. A similar procedure based on the random initial subset and the volume of the ellipsoid of Atkinson (1994), the outlier measure and the benchmark aforementioned is proposed in Chapter 5 to show the effectiveness of the proposed procedure when the initial basic subset contains outliers. This modified procedure is similar to the procedure proposed in Chapter 3 but the initial basic subset of the proposed procedure in Chapter 3 is replaced by a random initial subset which may include outliers. Moreover, the volume of the ellipsoid is used as an indicator of the best result as described in Chapter 5. The tables in

Chapter 5 show that the modified procedure still works even if the initial basic subset contains outliers. Chapter 6 discusses some alternative aspects of the proposed procedure, such as the alternative forms of the outlier measures in ordering the observations and an alternative way of finding the sample robust covariance matrix. Moreover, the reason that leads to the unsatisfactory performance of Method I and possible ways for improvement are described in Chapter 6. To conclude, the procedure proposed works nicely in identifying multiple multivariate outliers.

Appendix

The notations used in Tables 1 to 4 are described in the following:

n	:	sample size
p	:	dimension
μ	:	amount of shift
d	:	constant defining the amount of shift
ε	:	fraction of contamination
AMR	:	average misclassification rate
ASR	:	average success rate without misclassification

Note : d is shown in the table to show whether the constructed outliers are far outliers or close outliers.

Table 1: Effect of dimension p with n, d and ε fixed (a) sample covariance matrix, (b) robust estimate

				(a)		(b)	
n	p	d	ε	AMR	ASR	AMR	ASR
100	5	2	0.2	0.000000	**0.997000**	0.000000	**0.998000**
100	10	2	0.2	0.002500	**0.943000**	0.000000	**0.998000**
100	20	2	0.2	0.039750	**0.361000**	0.033500	**0.482000**
100	5	2	0.3	0.043714	0.404667	0.035143	0.436667
100	10	2	0.3	0.048857	0.227333	0.035429	0.358667
100	20	2	0.3	0.058571	0.110000	0.056000	0.103333
100	5	2	0.4	0.080333	0.089500	0.067667	0.115000
100	10	2	0.4	0.063333	0.080500	0.056000	0.085500
100	20	2	0.4	0.057000	0.077500	0.058667	0.073500
100	5	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
100	10	4	0.2	0.001000	**0.981000**	0.000000	**1.000000**
100	20	4	0.2	0.039500	**0.410000**	0.031250	**0.546000**
100	5	4	0.3	0.001429	**0.982667**	0.000000	**1.000000**
100	10	4	0.3	0.034857	**0.473333**	0.005429	**0.895333**
100	20	4	0.3	0.057429	**0.087333**	0.055714	**0.131333**
100	5	4	0.4	0.072000	0.160000	0.053000	0.243500
100	10	4	0.4	0.063333	0.079000	0.046667	0.180000
100	20	4	0.4	0.058000	0.074500	0.056000	0.081000
200	5	2	0.2	0.000000	**0.995500**	0.000000	**0.995500**
200	10	2	0.2	0.000000	**0.998500**	0.000000	**0.998500**
200	20	2	0.2	0.006750	**0.718000**	0.002875	**0.885500**
200	5	2	0.3	0.060143	0.176333	0.050714	0.229000
200	10	2	0.3	0.034857	0.120667	0.028857	0.166667
200	20	2	0.3	0.020143	0.090000	0.017000	0.109750
200	5	2	0.4	0.071833	0.095750	0.058000	0.109750
200	10	2	0.4	0.040167	0.050000	0.035833	0.055250
200	20	2	0.4	0.022833	0.027500	0.020000	0.025250
200	5	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
200	10	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
200	20	4	0.2	0.005875	**0.756500**	0.002125	**0.924000**
200	5	4	0.3	0.000000	**1.000000**	0.000000	**1.000000**
200	10	4	0.3	0.009429	**0.737667**	0.000000	**1.000000**
200	20	4	0.3	0.019714	**0.111000**	0.014143	**0.288667**
200	5	4	0.4	0.070167	0.094000	0.055667	0.117250
200	10	4	0.4	0.039833	0.049250	0.029500	0.133750
200	20	4	0.4	0.022333	0.026750	0.019833	0.029750

Table 1 (continue): Effect of dimension p with n, d and ε fixed (a) sample covariance matrix, (b) robust estimate

				(a)		(b)	
n	p	d	ε	AMR	ASR	AMR	ASR
400	5	2	0.2	0.000000	**0.997000**	0.000000	**0.997000**
400	10	2	0.2	0.000000	**0.998500**	0.000000	**0.998500**
400	20	2	0.2	0.000875	**0.921000**	0.000063	**0.980000**
400	5	2	0.3	0.052429	0.164333	0.045714	0.214667
400	10	2	0.3	0.028357	0.076167	0.129833	0.068167
400	20	2	0.3	0.008929	0.036500	0.008500	0.019000
400	5	2	0.4	0.064667	0.080250	0.054250	0.103625
400	10	2	0.4	0.031500	0.042500	0.028917	0.047500
400	20	2	0.4	0.010000	0.011875	0.009500	0.013750
400	5	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
400	10	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
400	20	4	0.2	0.001250	**0.901750**	0.000000	**1.000000**
400	5	4	0.3	0.000000	**0.999833**	0.000000	**0.999833**
400	10	4	0.3	0.001286	**0.961833**	0.000000	**1.000000**
400	20	4	0.3	0.007571	**0.213833**	0.003929	**0.626333**
400	5	4	0.4	0.063583	0.079500	0.050500	0.110125
400	10	4	0.4	0.030667	0.062000	0.025083	0.072750
400	20	4	0.4	0.010000	0.012000	0.008417	0.034000
800	5	2	0.2	0.000000	**0.997375**	0.000000	**0.997375**
800	10	2	0.2	0.000000	**0.999125**	0.000000	**0.999125**
800	20	2	0.2	0.000000	**0.999875**	0.000000	**0.999875**
800	5	2	0.3	0.055929	0.114667	0.047643	0.148583
800	10	2	0.3	0.023964	0.046083	0.021893	0.056083
800	20	2	0.3	0.005750	0.010000	0.005464	0.012250
800	5	2	0.4	0.062000	0.083500	0.051917	0.105500
800	10	2	0.4	0.026083	0.035625	0.023042	0.041125
800	20	2	0.4	0.006375	0.007438	0.006167	0.008625
800	5	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
800	10	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
800	20	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
800	5	4	0.3	0.000000	**0.999917**	0.000000	**0.999917**
800	10	4	0.3	0.000000	**1.000000**	0.000000	**1.000000**
800	20	4	0.3	0.003500	**0.484667**	0.000286	**0.940583**
800	5	4	0.4	0.061042	0.084375	0.048750	0.113125
800	10	4	0.4	0.026083	0.035188	0.021083	0.045188
800	20	4	0.4	0.006333	0.007312	0.005708	0.009750

Table 2: Effect of sample size n with d, p and ε fixed (a) sample covariance matrix, (b) robust estimate

				(a)		(b)	
n	p	d	ε	AMR	ASR	AMR	ASR
100	5	2	0.2	0.000000	**0.997000**	0.000000	**0.998000**
200	5	2	0.2	0.000000	**0.995500**	0.000000	**0.995500**
400	5	2	0.2	0.000000	**0.997000**	0.000000	**0.997000**
800	5	2	0.2	0.000000	**0.997375**	0.000000	**0.997375**
100	5	2	0.3	0.043714	0.404667	0.035143	0.436667
200	5	2	0.3	0.060143	0.176333	0.050714	0.229000
400	5	2	0.3	0.052429	0.164333	0.045714	0.214667
800	5	2	0.3	0.055929	0.114667	0.047643	0.148583
100	5	2	0.4	0.080333	0.089500	0.067667	0.115000
200	5	2	0.4	0.071833	0.095750	0.058000	0.109750
400	5	2	0.4	0.064667	0.080250	0.054250	0.103625
800	5	2	0.4	0.062000	0.083500	0.051917	0.105500
100	10	2	0.2	0.002500	**0.943000**	0.000000	**0.998000**
200	10	2	0.2	0.000000	**0.998500**	0.000000	**0.998500**
400	10	2	0.2	0.000000	**0.998500**	0.000000	**0.998500**
800	10	2	0.2	0.000000	**0.999125**	0.000000	**0.999125**
100	10	2	0.3	0.048857	0.227333	0.035429	0.358667
200	10	2	0.3	0.034857	0.120667	0.028857	0.166667
400	10	2	0.3	0.028357	0.076167	0.129833	0.068167
800	10	2	0.3	0.023964	0.046083	0.021893	0.056083
100	10	2	0.4	0.063333	0.080500	0.056000	0.085500
200	10	2	0.4	0.040167	0.050000	0.035833	0.055250
400	10	2	0.4	0.031500	0.042500	0.028917	0.047500
800	10	2	0.4	0.026083	0.035625	0.023042	0.041125
100	20	2	0.2	0.039750	**0.361000**	0.033500	**0.482000**
200	20	2	0.2	0.006750	**0.718000**	0.002875	**0.885500**
400	20	2	0.2	0.000875	**0.921000**	0.000063	**0.980000**
800	20	2	0.2	0.000000	**0.999875**	0.000000	**0.999875**
100	20	2	0.3	0.058571	0.110000	0.056000	0.103333
200	20	2	0.3	0.020143	0.090000	0.017000	0.109750
400	20	2	0.3	0.008929	0.036500	0.008500	0.019000
800	20	2	0.3	0.005750	0.010000	0.005464	0.012250
100	20	2	0.4	0.057000	0.077500	0.058667	0.073500
200	20	2	0.4	0.022833	0.027500	0.020000	0.025250
400	20	2	0.4	0.010000	0.011875	0.009500	0.013750
800	20	2	0.4	0.006375	0.007438	0.006167	0.008625

Table 2 (continue): Effect of sample size n with d, p and ε fixed (a) sample covariance matrix, (b) robust estimate

				(a)		(b)	
n	p	d	ε	AMR	ASR	AMR	ASR
100	5	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
200	5	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
400	5	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
800	5	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
100	5	4	0.3	0.001429	**0.982667**	0.000000	**1.000000**
200	5	4	0.3	0.000000	**1.000000**	0.000000	**1.000000**
400	5	4	0.3	0.000000	**0.999833**	0.000000	**0.999833**
800	5	4	0.3	0.000000	**0.999917**	0.000000	**0.999917**
100	5	4	0.4	0.072000	0.160000	0.053000	0.243500
200	5	4	0.4	0.070167	0.094000	0.055667	0.117250
400	5	4	0.4	0.063583	0.079500	0.050500	0.110125
800	5	4	0.4	0.061042	0.084375	0.048750	0.113125
100	10	4	0.2	0.001000	**0.981000**	0.000000	**1.000000**
200	10	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
400	10	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
800	10	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
100	10	4	0.3	0.034857	**0.473333**	0.005429	**0.895333**
200	10	4	0.3	0.009429	**0.737667**	0.000000	**1.000000**
400	10	4	0.3	0.001286	**0.961833**	0.000000	**1.000000**
800	10	4	0.3	0.000000	**1.000000**	0.000000	**1.000000**
100	10	4	0.4	0.063333	0.079000	0.046667	0.180000
200	10	4	0.4	0.039833	0.049250	0.029500	0.133750
400	10	4	0.4	0.030667	0.062000	0.025083	0.072750
800	10	4	0.4	0.026083	0.035188	0.021083	0.045188
100	20	4	0.2	0.039500	**0.410000**	0.031250	**0.546000**
200	20	4	0.2	0.005875	**0.756500**	0.002125	**0.924000**
400	20	4	0.2	0.001250	**0.901750**	0.000000	**1.000000**
800	20	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
100	20	4	0.3	0.057429	0.087333	0.055714	0.131333
200	20	4	0.3	0.019714	0.111000	0.014143	0.288667
400	20	4	0.3	0.007571	0.213833	0.003929	0.626333
800	20	4	0.3	0.003500	0.484667	0.000286	0.940583
100	20	4	0.4	0.058000	0.074500	0.056000	0.081000
200	20	4	0.4	0.022333	0.026750	0.019833	0.029750
400	20	4	0.4	0.010000	0.012000	0.008417	0.034000
800	20	4	0.4	0.006333	0.007312	0.005708	0.009750

Table 3: Effect of amount of shift μ with n, p and ε fixed (a) sample covariance matrix, (b) robust estimate

				(a)		(b)	
n	p	d	ε	AMR	ASR	AMR	ASR
100	5	2	0.2	0.000000	**0.997000**	0.000000	**0.998000**
100	5	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
100	5	2	0.3	0.043714	0.404667	0.035143	0.436667
100	5	4	0.3	0.001429	0.982667	0.000000	1.000000
100	5	2	0.4	0.080333	0.089500	0.067667	0.115000
100	5	4	0.4	0.072000	0.160000	0.053000	0.243500
100	10	2	0.2	0.002500	**0.943000**	0.000000	**0.998000**
100	10	4	0.2	0.001000	**0.981000**	0.000000	**1.000000**
100	10	2	0.3	0.048857	0.227333	0.035429	0.358667
100	10	4	0.3	0.034857	0.473333	0.005429	0.895333
100	10	2	0.4	0.063333	0.080500	0.056000	0.085500
100	10	4	0.4	0.063333	0.079000	0.046667	0.180000
100	20	2	0.2	0.039750	0.361000	0.033500	0.482000
100	20	4	0.2	0.039500	0.410000	0.031250	0.546000
100	20	2	0.3	0.058571	0.110000	0.056000	0.103333
100	20	4	0.3	0.057429	0.087333	0.055714	0.131333
100	20	2	0.4	0.057000	0.077500	0.058667	0.073500
100	20	4	0.4	0.058000	0.074500	0.056000	0.081000
200	5	2	0.2	0.000000	**0.995500**	0.000000	**0.995500**
200	5	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
200	5	2	0.3	0.060143	0.176333	0.050714	0.229000
200	5	4	0.3	0.000000	1.000000	0.000000	1.000000
200	5	2	0.4	0.071833	0.095750	0.058000	0.109750
200	5	4	0.4	0.070167	0.094000	0.055667	0.117250
200	10	2	0.2	0.000000	**0.998500**	0.000000	**0.998500**
200	10	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
200	10	2	0.3	0.034857	0.120667	0.028857	0.166667
200	10	4	0.3	0.009429	0.737667	0.000000	1.000000
200	10	2	0.4	0.040167	0.050000	0.035833	0.055250
200	10	4	0.4	0.039833	0.049250	0.029500	0.133750
200	20	2	0.2	0.006750	0.718000	0.002875	0.885500
200	20	4	0.2	0.005875	0.756500	0.002125	0.924000
200	20	2	0.3	0.020143	0.090000	0.017000	0.109750
200	20	4	0.3	0.019714	0.111000	0.014143	0.288667
200	20	2	0.4	0.022833	0.027500	0.020000	0.025250
200	20	4	0.4	0.022333	0.026750	0.019833	0.029750

Table 3 (continue): Effect of amount of shift μ with n, p and ε fixed (a) sample covariance matrix, (b) robust estimate

				(a)		(b)	
n	p	d	ε	AMR	ASR	AMR	ASR
400	5	2	0.2	0.000000	**0.997000**	0.000000	**0.997000**
400	5	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
400	5	2	0.3	0.052429	0.164333	0.045714	0.214667
400	5	4	0.3	0.000000	0.999833	0.000000	0.999833
400	5	2	0.4	0.064667	0.080250	0.054250	0.103625
400	5	4	0.4	0.063583	0.079500	0.050500	0.110125
400	10	2	0.2	0.000000	**0.998500**	0.000000	**0.998500**
400	10	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
400	10	2	0.3	0.028357	0.076167	0.129833	0.068167
400	10	4	0.3	0.001286	0.961833	0.000000	1.000000
400	10	2	0.4	0.031500	0.042500	0.028917	0.047500
400	10	4	0.4	0.030667	0.062000	0.025083	0.072750
400	20	2	0.2	0.000875	**0.921000**	0.000063	**0.980000**
400	20	4	0.2	0.001250	**0.901750**	0.000000	**1.000000**
400	20	2	0.3	0.008929	0.036500	0.008500	0.019000
400	20	4	0.3	0.007571	0.213833	0.003929	0.626333
400	20	2	0.4	0.010000	0.011875	0.009500	0.013750
400	20	4	0.4	0.010000	0.012000	0.008417	0.034000
800	5	2	0.2	0.000000	**0.997375**	0.000000	**0.997375**
800	5	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
800	5	2	0.3	0.055929	0.114667	0.047643	0.148583
800	5	4	0.3	0.000000	0.999917	0.000000	0.999917
800	5	2	0.4	0.062000	0.083500	0.051917	0.105500
800	5	4	0.4	0.061042	0.084375	0.048750	0.113125
800	10	2	0.2	0.000000	**0.999125**	0.000000	**0.999125**
800	10	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
800	10	2	0.3	0.023964	0.046083	0.021893	0.056083
800	10	4	0.3	0.000000	1.000000	0.000000	1.000000
800	10	2	0.4	0.026083	0.035625	0.023042	0.041125
800	10	4	0.4	0.026083	0.035188	0.021083	0.045188
800	20	2	0.2	0.000000	**0.999875**	0.000000	**0.999875**
800	20	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
800	20	2	0.3	0.005750	0.010000	0.005464	0.012250
800	20	4	0.3	0.003500	0.484667	0.000286	0.940583
800	20	2	0.4	0.006375	0.007438	0.006167	0.008625
800	20	4	0.4	0.006333	0.007312	0.005708	0.009750

Table 4: Effect of fraction of contamination ε with n, p and d fixed (a) sample covariance matrix, (b) robust estimate

				(a)		(b)	
n	p	d	ε	AMR	ASR	AMR	ASR
100	5	2	0.2	0.000000	0.997000	0.000000	0.998000
100	5	2	0.3	0.043714	0.404667	0.035143	0.436667
100	5	2	0.4	0.080333	0.089500	0.067667	0.115000
100	5	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
100	5	4	0.3	0.001429	**0.982667**	0.000000	**1.000000**
100	5	4	0.4	0.072000	**0.160000**	0.053000	**0.243500**
100	10	2	0.2	0.002500	0.943000	0.000000	0.998000
100	10	2	0.3	0.048857	0.227333	0.035429	0.358667
100	10	2	0.4	0.063333	0.080500	0.056000	0.085500
100	10	4	0.2	0.001000	**0.981000**	0.000000	**1.000000**
100	10	4	0.3	0.034857	**0.473333**	0.005429	**0.895333**
100	10	4	0.4	0.063333	**0.079000**	0.046667	**0.180000**
100	20	2	0.2	0.039750	0.361000	0.033500	0.482000
100	20	2	0.3	0.058571	0.110000	0.056000	0.103333
100	20	2	0.4	0.057000	0.077500	0.058667	0.073500
100	20	4	0.2	0.039500	0.410000	0.031250	0.546000
100	20	4	0.3	0.057429	0.087333	0.055714	0.131333
100	20	4	0.4	0.058000	0.074500	0.056000	0.081000
200	5	2	0.2	0.000000	0.995500	0.000000	0.995500
200	5	2	0.3	0.060143	0.176333	0.050714	0.229000
200	5	2	0.4	0.071833	0.095750	0.058000	0.109750
200	5	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
200	5	4	0.3	0.000000	**1.000000**	0.000000	**1.000000**
200	5	4	0.4	0.070167	**0.094000**	0.055667	**0.117250**
200	10	2	0.2	0.000000	0.998500	0.000000	0.998500
200	10	2	0.3	0.034857	0.120667	0.028857	0.166667
200	10	2	0.4	0.040167	0.050000	0.035833	0.055250
200	10	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
200	10	4	0.3	0.009429	**0.737667**	0.000000	**1.000000**
200	10	4	0.4	0.039833	**0.049250**	0.029500	**0.133750**
200	20	2	0.2	0.006750	0.718000	0.002875	0.885500
200	20	2	0.3	0.020143	0.090000	0.017000	0.109750
200	20	2	0.4	0.022833	0.027500	0.020000	0.025250
200	20	4	0.2	0.005875	0.756500	0.002125	0.924000
200	20	4	0.3	0.019714	0.111000	0.014143	0.288667
200	20	4	0.4	0.022333	0.026750	0.019833	0.029750

Table 4 (continue): Effect of fraction of contamination ε with n, p and d fixed (a) sample covariance matrix, (b) robust estimate

				(a)		(b)	
n	p	d	ε	AMR	ASR	AMR	ASR
400	5	2	0.2	0.000000	0.997000	0.000000	0.997000
400	5	2	0.3	0.052429	0.164333	0.045714	0.214667
400	5	2	0.4	0.064667	0.080250	0.054250	0.103625
400	5	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
400	5	4	0.3	0.000000	**0.999833**	0.000000	**0.999833**
400	5	4	0.4	0.063583	**0.079500**	0.050500	**0.110125**
400	10	2	0.2	0.000000	0.998500	0.000000	0.998500
400	10	2	0.3	0.028357	0.076167	0.129833	0.068167
400	10	2	0.4	0.031500	0.042500	0.028917	0.047500
400	10	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
400	10	4	0.3	0.001286	**0.961833**	0.000000	**1.000000**
400	10	4	0.4	0.030667	**0.062000**	0.025083	**0.072750**
400	20	2	0.2	0.000875	0.921000	0.000063	0.980000
400	20	2	0.3	0.008929	0.036500	0.008500	0.019000
400	20	2	0.4	0.010000	0.011875	0.009500	0.013750
400	20	4	0.2	0.001250	0.901750	0.000000	1.000000
400	20	4	0.3	0.007571	0.213833	0.003929	0.626333
400	20	4	0.4	0.010000	0.012000	0.008417	0.034000
800	5	2	0.2	0.000000	0.997375	0.000000	0.997375
800	5	2	0.3	0.055929	0.114667	0.047643	0.148583
800	5	2	0.4	0.062000	0.083500	0.051917	0.105500
800	5	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
800	5	4	0.3	0.000000	**0.999917**	0.000000	**0.999917**
800	5	4	0.4	0.061042	**0.084375**	0.048750	**0.113125**
800	10	2	0.2	0.000000	0.999125	0.000000	0.999125
800	10	2	0.3	0.023964	0.046083	0.021893	0.056083
800	10	2	0.4	0.026083	0.035625	0.023042	0.041125
800	10	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
800	10	4	0.3	0.000000	**1.000000**	0.000000	**1.000000**
800	10	4	0.4	0.026083	**0.035188**	0.021083	**0.045188**
800	20	2	0.2	0.000000	0.999875	0.000000	0.999875
800	20	2	0.3	0.005750	0.010000	0.005464	0.012250
800	20	2	0.4	0.006375	0.007438	0.006167	0.008625
800	20	4	0.2	0.000000	**1.000000**	0.000000	**1.000000**
800	20	4	0.3	0.003500	**0.484667**	0.000286	**0.940583**
800	20	4	0.4	0.006333	**0.007312**	0.005708	**0.009750**

References

Atkinson, A. C. (1986). Masking Unmasked. *Biometrika, 73*, 533–541.

Atkinson, A. C. (1994). Fast Very Robust Methods for the Detection of Multiple Outliers. *Journal of the American Statistical Association, 89*, 1329–1339.

Atkinson, A. C. and Mulira, H. M. (1993). The Stalactite Plot for the Detection of Multivariate Outliers. *Statistics and Computing, 3*, 27–35.

Brownlee, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering,* 2nd edn. New York: Wiley.

Cook, R. D. (1986). Assessment of Local Influence (with Discussion). *Journal of the Royal Statistical Society, B, 48*, 133–169.

Hadi, A. (1992). Identifying Multiple Outliers in Multivariate Data. *Journal of the Royal Statistical Society, B, 54*, 761–771.

Hadi, A. (1994). A Modification of a Method for the Detection of outliers in Multivariate Samples. *Journal of the Royal Statistical Society, B, 56*, 393–396.

Hawkins, D. M., Bradu, D. and Kass, G. V. (1984). Location of Several Outliers in Multiple Regression Data Using Elemental Subsets. *Technometrics, 26*, 197–208.

Lopuhaä, H. P. and Rousseeuw, P. J. (1991). Breakdown Points of Affine-Equivariant Estimators of Multivariate Location and Covariance Matrices. *The Annals of Statistics, 19*, 229–248.

Poon, W. Y., Lew, S. F. and Poon, Y. S. (2000). A Local Influence Approach to Identify Multiple Multivariate Outliers. *British Journal of Mathematical and Statistical Psychology*, to appear.

Poon, W. Y. and Poon, Y. S. (1999). Conformal Normal Curvature and Assessment of Local Influence. *Journal of the Royal Statistical Society, B, 61*, 51–61.

Rocke, D. M. and Woodruff, D. L. (1996). Identification of Outliers in Multivariate Data. *Journal of the American Statistical Association, 91*, 1047–1061.

Rousseeuw, P. J. and Leroy, A. (1987). *Robust Regression and Outlier Detection.* New York: Wiley.

Rousseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking Multivariate Outliers and Leverage Points. *Journal of the American Statistical Association, 85*, 633–639.